Université de Montréal

Analyse transcriptomique et applications en développement préclinique des médicaments

Par Nehme El-Hachem

Département de Sciences Biomédicales
Faculté de Médecine

Thèse présentée
en vue de l'obtention du grade de Doctorat
en Sciences Biomédicales

Août, 2016

Résumé

L'émergence des Mégadonnées (« Big Data ») en biologie moléculaire, surtout à travers la transcriptomique, a révolutionné la façon dont nous étudions diverses disciplines telles que le processus de développement du médicament ou la recherche sur le cancer. Ceci fut associé à un nouveau concept, la médecine de précision, dont le principal but est de comprendre les mécanismes moléculaires entraînant une meilleure réponse thérapeutique chez le patient.

Cette thèse est à mi-chemin entre les études pharmaco — et toxicogénomiques expérimentales, et les études cliniques et translationnelles. Le but de cette thèse est surtout de montrer le potentiel et les limites de ces jeux de données et leur pertinence pour la découverte de biomarqueurs de réponse ainsi que la compréhension des mécanismes d'action/toxicité de médicaments, en vue d'utiliser ces informations à des fins thérapeutiques. L'originalité de cette thèse réside dans son approche globale pour analyser les plus larges jeux de données pharmaco/toxicogénomiques publiés à ce jour et ceci pour : 1) Aborder la notion de biomarqueurs de réponse aux médicaments en pharmacogénomique du cancer, en étudiant les facteurs discordants entre deux grandes études publiées en 2012; 2) Comprendre le mécanisme d'action des médicaments et construire une taxonomie performante en utilisant une approche intégrative; et 3) Créer un répertoire toxicogénomique à partir des hépatocytes humains, exposés à différentes classes de médicaments et composés chimiques.

Mes contributions principales sont les suivantes :

- J'ai développé une approche bioinformatique pour étudier les facteurs discordants entre deux grandes études pharmacogénomiques et suggérées que les différences observées émergeaient plutôt de l'absence de standardisation des mesures pharmacologiques qui pourrait limiter la validation de biomarqueurs de réponse aux médicaments.
- J'ai implémenté une approche bioinformatique qui montre la supériorité de l'intégration tenant en compte des différents paramètres pour les médicaments (structure, cytotoxicité, perturbation du transcriptome) afin d'élucider leur mécanisme d'action (MoA).

- J’ai développé un pipeline bioinformatique pour étudier le niveau de conservation des mécanismes moléculaires entre les études toxicogénomiques *in vivo* et *in vitro* démontrant que les hépatocytes humains sont un modèle fiable pour détecter les produits toxiques hépatocarcinogènes.

Au total, nos études ont permis de fournir un cadre de travail original pour l’exploitation de différents types de données transcriptomiques pour comprendre l’impact des produits chimiques sur la biologie cellulaire.

Mots-clés : Transcriptomique, bioinformatique, médicaments, biomarqueurs de réponse, mécanisme d’action, toxicité, lignée cellulaire, microarrays, pharmaco-toxicogénomique

Abstract

The emergence of Big Data in molecular biology, especially through the study of transcriptomics, has revolutionized the way we look at various disciplines, such as drug development and cancer research. Big data analysis is an important part of the concept of precision medicine, which primary purpose is to understand the molecular mechanisms leading to better therapeutic response in patients.

This thesis is halfway between pharmaco-toxicogenomics experimental studies, and clinical and translational studies. The aim of this thesis is mainly to show the potential and limitations of these studies and their relevance, especially for the discovery of drug response biomarkers and understanding the drug mechanisms (targets, toxicities). This thesis is an original work since it proposes a global approach to analyzing the largest pharmaco-toxicogenomic datasets available to date. The key aims were: 1) Addressing the challenge of reproducibility for biomarker discovery in cancer pharmacogenomics, by comparing two large pharmacogenomics studies published in 2012; 2) Understanding drugs mechanism of action using an integrative approach to generate a superior drug-taxonomy; and 3) Evaluating the conservation of toxicogenomic responses in primary hepatocytes vs. in vivo liver samples in order to check the feasability of cell models in toxicology studies.

My main contributions can be summarized as follow:

- I developed a bioinformatics pipeline to study the factors that trigger (in)consistency between two major pharmacogenomic studies. I suggested that the observed differences emerged from the non-standardization of pharmacological measurements, which could limit the validation of drug response biomarker.
- I implemented a bioinformatics pipeline that demonstrated the superiority of the integrative approach, since it takes into account different parameters for the drug (structure, cytotoxicity, transcriptional perturbation) to elucidate the mechanism of action (MoA).
- I developed a bioinformatics pipeline to study the level of conservation of toxicity mechanisms between the in vivo and in vitro system, showing that human hepatocytes is a reliable model for hepatocarcinogens testing.

Overall, our studies have provided a unique framework to leverage various types of transcriptomic data in order to understand the impact of chemicals on cell biology.

Keywords : Transcriptomics, bioinformatics, chemical compounds, response biomarkers, mechanism of action, toxicity, cell lines, microarrays, pharmaco-toxicogenomics

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles

ADME absorption-distribution-metabolims-excretion

ADN acide désoxyribonucléique

ADNc acide désoxyribonucléique complementaire

APC affinity propagation clustering

ARNm acide ribonucléique messager codant

ATC anatomical therapeutic classification

AUC area under the curve

CCLE cancer cell line encyclopedia

CDF Chip Description File

CDK cyclin dependant kinase

CMAP connectivity map

CNV copy number variant

CTRP cancer therapeutics response portal

DNF drug network fusion

FDR false discovery rate

GCOS GeneChip® Operating Software

GDSC Genomics of Drug Sensitivity in Cancer

GEO gene expression omnibus

GO gene ontology

GSEA gene set enrichment analysis

HDACs histone deacetylase

IC_{50} inhibitory concentration of 50%

ISA Iterative Signature Algorithm

LIMMA linear models for microarray data

LINCS Library of Integrated Cellular Signatures

MAS5 Microarray affymetrix suite

MM mismatch probe

MoA mechanism of action

NAMPT Nicotinamide phosphoribosyltransferase

NCI-60 national cancer institute-60 cell lines

NES normalized enrichment score

NSAIDs non steroidal antiinflamatory drugs

PARP poly (ADP-ribose) polymerase

PHH primary human hepatocytes

PM perfect match probe

PRH primary rat hepatocytes

QSAR quantitative structure activity relationship

REACH Registration, Evaluation, Authorization and restriction of Chemicals

RLV rat liver

RMA robust multiarray average

RNA-seq high throughput RNA sequencing

RNAi RNA interference

ROC Receiver Operating Characteristic curves

SMILES Simplified Molecular Input Line Entry Specification

SNF similarity network fusion

TGF-βR Transforming growth factor beta receptors

À *mes parents,*

Remerciements

Je désire premièrement remercier la direction de l'Institut de recherches cliniques de Montréal (IRCM) qui m'a accueilli pendant ces 4 années de doctorat. Je voudrais remercier tout spécialement mon directeur de recherche Dr Benjamin Haibe-Kains pour sa générosité, son aide et sa disponibilité tout au long de ma thèse malgré son départ pour Toronto. Il est resté un excellent mentor et a nourri chez moi la curiosité scientifique surtout pour la génomique computationnelle. Il était omniprésent et me réconforta même dans les moments les plus décourageants. Dr Haibe-Kains, je vous souhaite beaucoup de succès et d'énergie.

Je remercie mon codirecteur Jacques Archambault, avec qui j'ai eu la chance de collaborer et qui a accepté de me superviser à l'IRCM, pour sa générosité, son aide, ses encouragements ainsi que pour les fréquentes discussions scientifiques enrichissantes. Dr Archambault, je vous souhaite une excellente carrière à l'université McGill.

Je veux aussi remercier mes autres collaborateurs à l'Université de Harvard : Dr Hugo Aerts et Patrick Grossmann, avec qui j'ai partagé l'expertise en toxicogénomique.

Je remercie tous mes collègues à l'IRCM et aussi ceux de Toronto même si on s'est vu quelques fois seulement. Merci pour tous les projets collaboratifs et bonne continuation!

Je remercie les membres du jury pour l'intérêt porté à ce travail en acceptant de réviser cette thèse.

J'aimerais remercier la communauté de l'Institut de recherches cliniques de Montréal (IRCM) et particulièrement Virginie Leduc pour toute son aide avec toute la paperasse administrative. Je voudrais aussi remercier le département de Sciences biomédicales qui a pris soin de mon dossier étudiant.

Merci aussi à mes parents qui sont toujours fiers de moi. Un gros merci à ma conjointe Carole qui m'a encouragée durant la rédaction de cette thèse. Son gros sourire le matin, son énergie et dynamisme contagieux m'ont permis de franchir toutes les barrières difficiles durant ses 4 années de thèse.

Enfin, à ma fille Grace-Catherine, je souhaite que tes yeux soient toujours pleins de cette curiosité et cette vivacité, et que tu me regardes toujours comme tu le fais maintenant; J'espère mériter ton amour et que tes sourires seront toujours aussi réels et beaux. papa :)

CHAPITRE 1 : INTRODUCTION

1.1 Mégadonnées (« Big Data »), développement de médicaments et Médecine de précision

Le développement de médicaments a toujours été un processus long et coûteux. Au cours de la dernière décennie, la génération et l'analyse de mégadonnées « Big Data » a progressé à un rythme sans précédent basé sur le développement de larges bases de données pharmacogénomiques, chemo-génomiques et toxicogénomiques. Ce changement de paradigme a permis l'identification de nouveaux médicaments spécifiques pour certains types de cancers ou le repositionnement de médicaments approuvés ou en cours de développement pour d'autres indications cliniques (Kim, Goossens, & Hoshida, 2016).

.

Cependant, le développement de médicaments fait face à plusieurs défis majeurs :

1— Une capacité limitée à comprendre et caractériser les contextes biologiques divers surtout dans des maladies complexes telles que le cancer. En effet, il existe une nécessité à identifier les aberrations génomiques ou les variations transcriptomiques prédictives de la réponse à un médicament (Barretina et al., 2012; Garnett et al., 2012).

2— Les méthodes traditionnelles comptent sur une vision simplifiée de la biologie étant limitée à la manipulation d'une seule cible ou d'une voie moléculaire. Cependant, la communauté scientifique en est venue à mieux comprendre l'effet polypharmacologique des médicaments en intégrant les données chimiques, transcriptomiques et pharmacologiques, pour mieux caractériser les voies de signalisations moléculaires perturbées par le médicament (Chen et al., 2015; Iorio et al., 2010).

3— Un manque de modèle expérimental adéquat pour tester la toxicité des médicaments chez l'Humain. Il sera donc important d'exploiter les données toxicogénomiques dans un modèle cellulaire hépatique humain afin d'identifier des voies de signalisation responsables de la toxicité/carcinogénicité des médicaments. Ceci pourrait avantageusement remplacer les modèles de rongeurs moins fiables et coûteux (Grinberg et al., 2014).

L'émergence des mégadonnées en biologie a fondamentalement révolutionné la façon dont nous étudions la biologie moléculaire et le développement de médicaments. Nous sommes

maintenant au point de résoudre certains des obstacles mentionnés ci-dessus et de faciliter la compréhension des mécanismes d'action des médicaments pour une meilleure utilisation dans la pratique clinique. L'accumulation de ces données, couvrant une variété de contexte cellulaire/systèmes biologiques, est en train de devenir une ressource fiable pour démontrer la preuve de concept dans le cadre de la médecine de précision (Figure 1.1).

Cette thèse sera consacrée à l'analyse des mégadonnées générées à partir des puces à ADN (expression microarrays), une technologie largement appliquée pour la mesure des copies d'ARNm dans les tissus/cellules à l'échelle génomique.

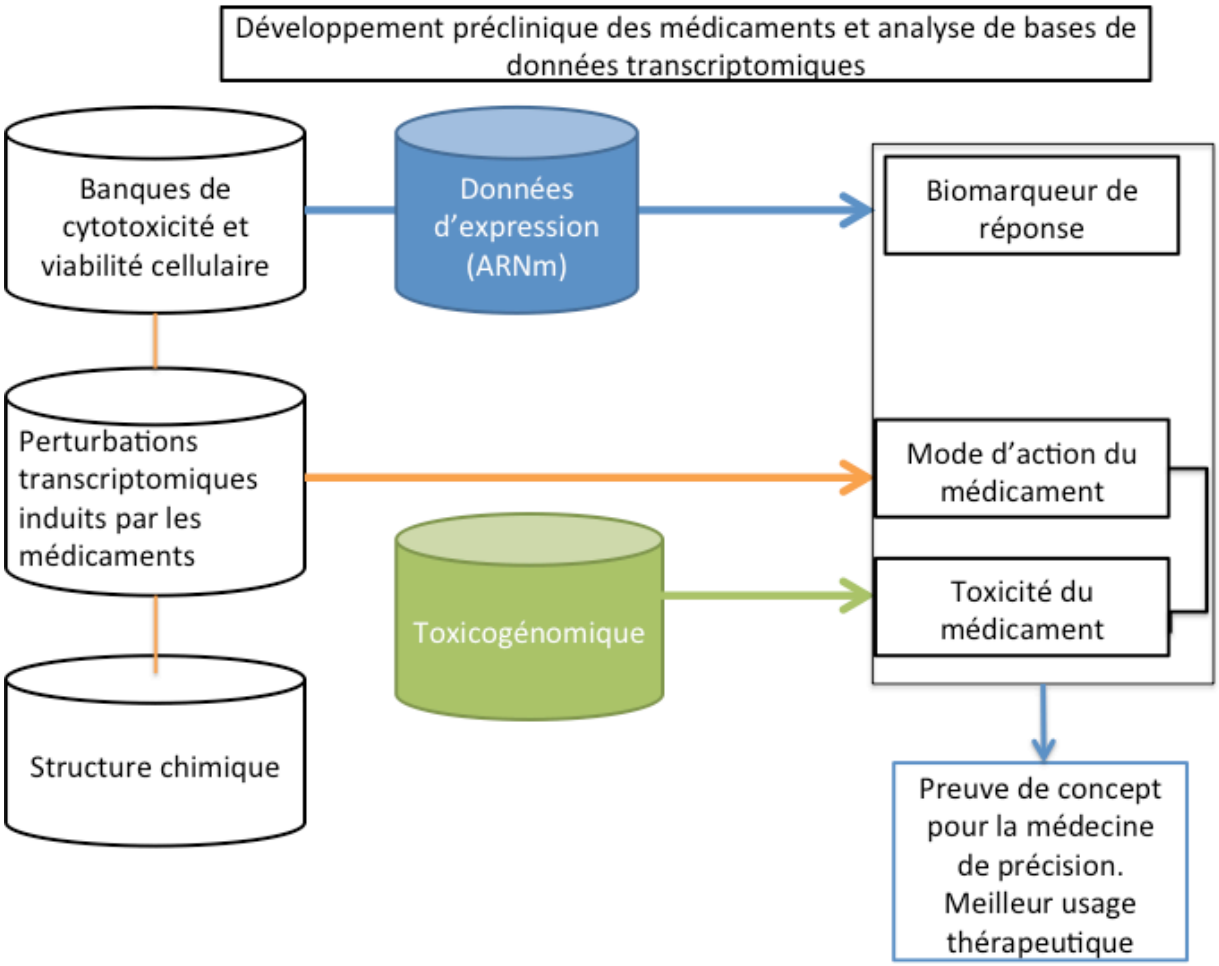

Figure 1.1 Cette figure montre une vue d'ensemble du projet de thèse ainsi que les différents types de données transcriptomiques abordés. Le cadre de travail montre comment l'analyse des données transcriptomiques et pharmacologiques à partir de modèles cellulaires pourra améliorer la compréhension du mécanisme d'action des médicaments dans les phases précoces du développement préclinique du médicament et ceci dans le but de prédire une meilleure efficacité dans les études translationnelles. Ceci n'est pas une figure générique, mais justement adaptée au projet de doctorat.

1.1.1 Microarrays et applications

L'utilisation de puces à ADN pour mesurer l'expression génique est particulièrement importante dans le cancer (Sotiriou & Pusztai, 2009). En effet, l'accumulation des effets des anomalies cellulaires qui conduisent à l'apparition et la progression maligne est le résultat de plusieurs altérations au niveau de la séquence ou l'expression de gènes. Ces anomalies (épi)

génomiques, qui peuvent être héréditaires ou acquises, conduisent aux principales caractéristiques du cancer, à savoir : l'autosuffisance en signaux de croissance; l'insensibilité aux signaux inhibiteurs de prolifération; l'évasion de l'apoptose; l'acquisition du potentiel réplicatif non contrôlé; l'induction de l'angiogenèse; l'invasion et les métastases (Hanahan & Weinberg, 2011).

Comme on le verra dans les paragraphes suivants, l'utilisation de micropuces est pertinente pour comprendre la biologie du cancer et pour choisir l'option thérapeutique correspondante. Des applications importantes de cette technologie comprennent : la découverte d'indicateurs diagnostiques ou pronostiques ainsi que des biomarqueurs de la réponse au traitement; l'identification des gènes impliqués dans la sensibilité/résistance aux médicaments; l'identification et la validation de nouvelles cibles moléculaires et une meilleure compréhension du mode d'action moléculaire des médicaments; la prévision des effets toxiques potentiels au cours des études de développement précliniques; et enfin la sélection des patients les plus susceptibles de bénéficier du médicament et de l'utiliser dans les études pharmacogénomiques (Clarke, te Poele, Wooster, & Workman, 2001).

1.1.2 Du gène à la protéine

Les protéines (enzymes, récepteurs) sont les principaux éléments actifs des cellules. Ils remplissent de nombreuses fonctions clés et contribuent au maintien des cellules et des tissus. L'information pour produire les protéines nécessaires dans une cellule sous une condition particulière est contenue dans l'acide désoxyribonucléique (ADN). La séquence complète d'ADN (le génome), est organisée dans les chromosomes et les gènes. Le dogme central de la biologie moléculaire (Crick, 1958) décrit les deux principales étapes de la production de protéines à partir d'ADN. Durant la première étape ou transcription, l'acide ribonucléique (ARNm) est transcrit de l'ADN et, dans la deuxième étape, connue sous le nom de traduction, les protéines sont produites sur la base des informations provenant de l'ARNm.

Contrairement à l'ADN, qui demeure relativement stable au cours de la durée de vie d'un organisme, les niveaux d'ARNm varient au fil du temps et selon les types cellulaires. Cette variation est aussi observée dans des cellules sous des conditions différentes. Par exemple, les copies d'ARNm transcrites à partir d'un gène peuvent différer de celles transcrites à partir du

même gène dans une pathologie tel que le cancer. Dans ce cas, ce gène est exprimé de façon différentielle entre les deux conditions (normal et pathologique) (Alberts et al., 2014).

1.1.3 limites et défis computationnels

Le progrès des technologies de séquençage de deuxième génération (Quail et al. 2012) ont abouti à la génération de grandes quantités de données de séquences. Par conséquent, la biologie moderne présente maintenant de nouveaux défis en termes de gestion des données, de requêtes et d'analyses. L'ADN humain est composé d'environ 3 milliards de paires de base avec un génome représentant approximativement 100 gigaoctets (Go) de données. En 2011, la capacité de séquençage annuelle mondiale était estimée à 13 quadrillions de base (13×10^{15}) (Pollack 2011). La plupart des programmes d'analyse bio-informatiques sont complexes à installer et configurer et à entretenir, principalement parce qu'elles sont, pour la plupart, écrites par des scientifiques qui manquent de temps et parfois d'expertise pour écrire des codes modulables et bien documentés. En conséquence, l'analyse des mégas donnés nécessite un niveau avancé d'expertise technique. Actuellement, une question plus importante que le simple stockage des mégas donnés est de les traiter en temps opportun, et ensuite d'analyser ces données pour des fins biomédicales. Par exemple, MapReduce /Hadoop est une technologie de traitement et d'analyse de données qui a été révolutionnaire dans le domaine de l'informatique et est l'une des technologies les plus prometteuses pour gérer les mégas donnés (O'Driscoll et al. 2013). Un autre exemple est celui de l'utilisation de l'infonuagique tel que Cloud BioLinux (Krampis et al. 2012), développé au centre J. Craig Venter (JCVI). Il s'agit d'une machine virtuelle accessible au public qui est stockée sur Amazon EC2, basée sur une distribution Ubuntu Linux avec plus de 100 outils bioinformatiques préinstallés (Giardine et al. 2005). Il existe également des inconvénients associés à l'utilisation de l'infonuagique. Étant donné l'ampleur des données génomiques générées, l'un des défis les plus importants est la transmission de ces données sur Internet qui demandent souvent d'énormes ressources en bande passante et de temps de transferts. Par exemple, le Beijing Genomics Institute (BGI), l'un des principaux instituts mondiaux de recherche en génomique produit 200 génomes par jour, qui sont transportés manuellement via FedEx (Clarke et al. 2012). Aussi, la capacité de protéger les données médicales et génomiques est un défi de plus en plus important. L'infonuagique est une méthode performante pour gérer les mégas donnés, cependant, le

séquençage clinique doit répondre à des exigences réglementaires extrêmement rigoureuses, qui ne permettent pas toujours le transfert de données sur des infrastructures distantes (Schadt 2012)

La création de bases de données et leur interface pour des méga donnés ne correspondent pas à mon champ d'expertise. Cependant, certains membres du labo ont développé PharmacoDb, une application internet qui permet de cataloguer et chercher des médicaments ou lignées cellulaires d'intérêt parmi les données disponibles dans notre plateforme PharmacoGx. SQL a été utilisé pour les données pharmacologiques et NoSQL est considéré pour le stockage et indexage des données issues du profilage moléculaire à haut débit. Il est cependant trop tôt pour dire si cette approche sera appliquée avec succès. Une revue de littérature récente présente en détails comment ce type de base de données aide à gérer les données de séquençage d'ADN et la gestion des dossiers médicaux électroniques (Luo et al. 2016).

Il est important de noter que les banques de données pharmacogénomiques citées dans ma thèse sont publiques et ont été téléchargées pour un stockage local dans le laboratoire du Dr. Haibe-Kains à Toronto. La plupart des analyses ont nécessité des ressources computationnelles plus ou moins modestes (~100GB de mémoire vive et <300GB de stockage), car il s'agit de données de micropuces. Les banques de données originales ont été collectionnées, organisées, manuellement inspectées et standardisées. Plusieurs membres du laboratoire du Dr. Haibe-Kains ont contribué à ce travail qui est disponible au public sur le site de bioconductor https://bioconductor.org/packages/release/bioc/html/PharmacoGx.html

1.2 Mesure de l'expression des gènes et micropuces à ADN

1.2.1 La technologie Affymetrix

La technologie Affymetrix GeneChip utilise une seule puce (« microarray ») par échantillon pour mesurer l'abondance des ARNm. Chaque gène est représenté par un ensemble de sondes (« probe set »), avec ~40,000 ensembles de sondes sur une seule puce. Chaque ensemble de sondes se compose de 10-16 paires de sondes d'ADN. Pour chaque paire de sondes, il existe une sonde concordante (« Perfect match ou PM ») et une disconcordante (« miss match ou MM ») (Figure 1.2). La longueur de la séquence de chaque sonde est de 25 nucléotides. Les

sondes PM et MM ont des séquences de base identiques, sauf pour le milieu de la sonde; dans la sonde MM la base nucléique est complémentaire à celle de la sonde PM. Les sondes MM sont donc conçues pour mesurer l'intensité non spécifique (Dalma-Weiszhausz, Warrington, Tanimoto, & Miyada, 2006; Irizarry, Bolstad, et al., 2003)

Les sondes candidates pour chaque gène sont choisies à partir d'une séquence de référence représentant le gène. Un certain nombre de procédures sont utilisées pour estimer la spécificité et la sensibilité de chaque sonde, dans le but de minimiser le risque d'hybridation non spécifique. En outre, la sélection de la sonde utilisée par Affymetrix est basée sur l'extrémité 3' de la séquence de référence à cause d'une présomption que la sonde oligo-T sera utilisé pour sélectionner les populations d'ARNm, et aussi parce que la divergence de séquence est généralement supérieure dans cette région (Mei et al., 2003)

.

Afin de mener une expérience de micropuces pour mesurer l'expression génique (génome codant), il faudra extraire les ARNm de la cellule et produire la banque ADNc (ADN complémentaire) correspondante. Dans cette étape, la transcriptase inverse synthétise un brin d'ADN qui correspond à la séquence de la molécule d'ARN. La raison est que la molécule d'ARN est moins stable que l'ADN, de sorte que la conversion contribue à maintenir la stabilité sans une perte de l'information génétique.

Dans l'étape suivante, l'échantillon d'ADNc doit être marqué par fluorescence. L'échantillon d'ADNc marqué est hybridé sur la micropuce à ADN et sur une base de complémentarité de séquence. Les molécules d'ADNc peuvent être identifiées sur la base de leurs fluorescences. L'intensité du signal est fonction de la quantité d'ADN hybridée, par conséquent ceci représente le niveau d'expression du gène correspondant (Kohane, Butte, & Kho, 2002)

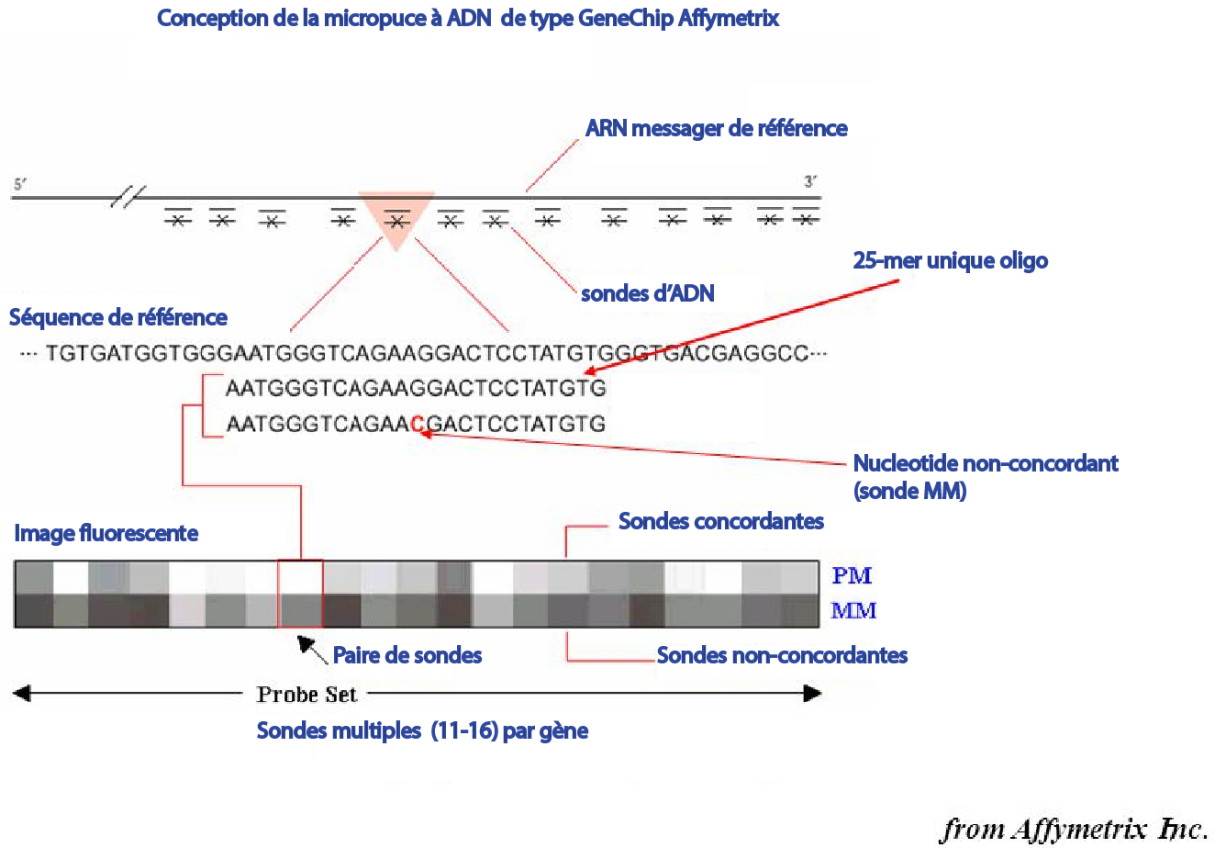

Figure 1.2 cette figure montre la conception de la sonde Affymetrix pour l'étude de l'expression des gènes (Affymetrix Inc.)

1.2.2 Aperçu de l'analyse des données de micropuces

Les informations concernant les ensembles de sondes sont disponibles dans ce qu'on appelle Chip Description File (CDF). Cependant, les annotations du transcriptome changent au fil du temps. Les CDFs sont rarement mis à jour, par conséquent, beaucoup de gens préfèrent utiliser des versions personnalisées de CDFs (Dai et al., 2005) qui représentent les annotations récentes au niveau du transcriptome. Dans ces fichiers, des sondes qui correspondent à plusieurs gènes sont filtrées et le reste d'entre eux sont regroupés par gène unique correspondant à la dernière version du génome. Les sondes peuvent avoir différentes intensités dépendamment du contexte technique et biologique.

La première étape dans l'analyse computationnelle des puces à ADN est de cribler l'ensemble des sondes sur la puce (fichier. CEL généré par la suite logiciel GCOS Affymetrix), ensuite obtenir une mesure d'expression unique pour chaque gène à travers l'ensemble des sondes. De nombreux algorithmes ont été mis au point pour cette procédure, le plus populaire d'entre eux étant RMA (Irizarry et al., 2003a) et MAS5 (Hubbell, Liu, & Mei, 2002). RMA est un logiciel rendu public http://bioconductor.org/packages/release/bioc/html/affy.html, alors que MAS5 est la propriété d'Affymetrix. La principale différence entre RMA et MAS5 est que RMA utilise seulement les valeurs de PM tandis que MAS5 utilise à la fois les valeurs de PM et MM. Dans l'algorithme MAS5, les valeurs MM sont soustraites des valeurs de PM pour tenir en compte l'hybridation non spécifique alors que RMA ignore les valeurs de MM. Aussi, MAS5 normalise indépendamment chaque microarray tandis que RMA normalise l'ensemble en même temps. Nous utilisons RMA dans les études décrites dans cette thèse, car cet algorithme permet de mieux corriger les biais expérimentaux en les modélisant au travers de tous les échantillons d'un ensemble de données d'intérêt (Irizarry et al., 2003b).

Une des principales utilisations des données des puces est l'identification des gènes différentiellement exprimés. Il existe de nombreux outils pour cette tâche, mais l'un des plus populaires est Limma (Smyth, 2004). Limma utilise un test t de Student modéré pour calculer les valeurs p (p-value). Une valeur de p peut-être interprétée comme la probabilité de faussement rejeter l'hypothèse nulle (erreur de type I). Par conséquent, la valeur p représente la significativité statistique des gènes exprimés dans deux conditions différentes. Limma calcule également les valeurs p ajustées qui sont plus communément connues sous le nom

False Discovery Rate (FDR). C'est une technique qui est utilisée pour tenir compte de tests multiples et qui désigne le pourcentage de faux positifs parmi les hypothèses testées (Benjamini & Hochberg, 1995). De plus, il existe des techniques statistiques plus avancées telles que GSEA (Subramanian et al., 2005) qui peuvent fournir des renseignements sur le rôle des gènes différentiellement exprimés dans le contrôle des voies biochimiques impliquées dans un type de cancer ou dans une condition spécifique (contrôle vs traitement avec un médicament par exemple) (Figure 1.3).

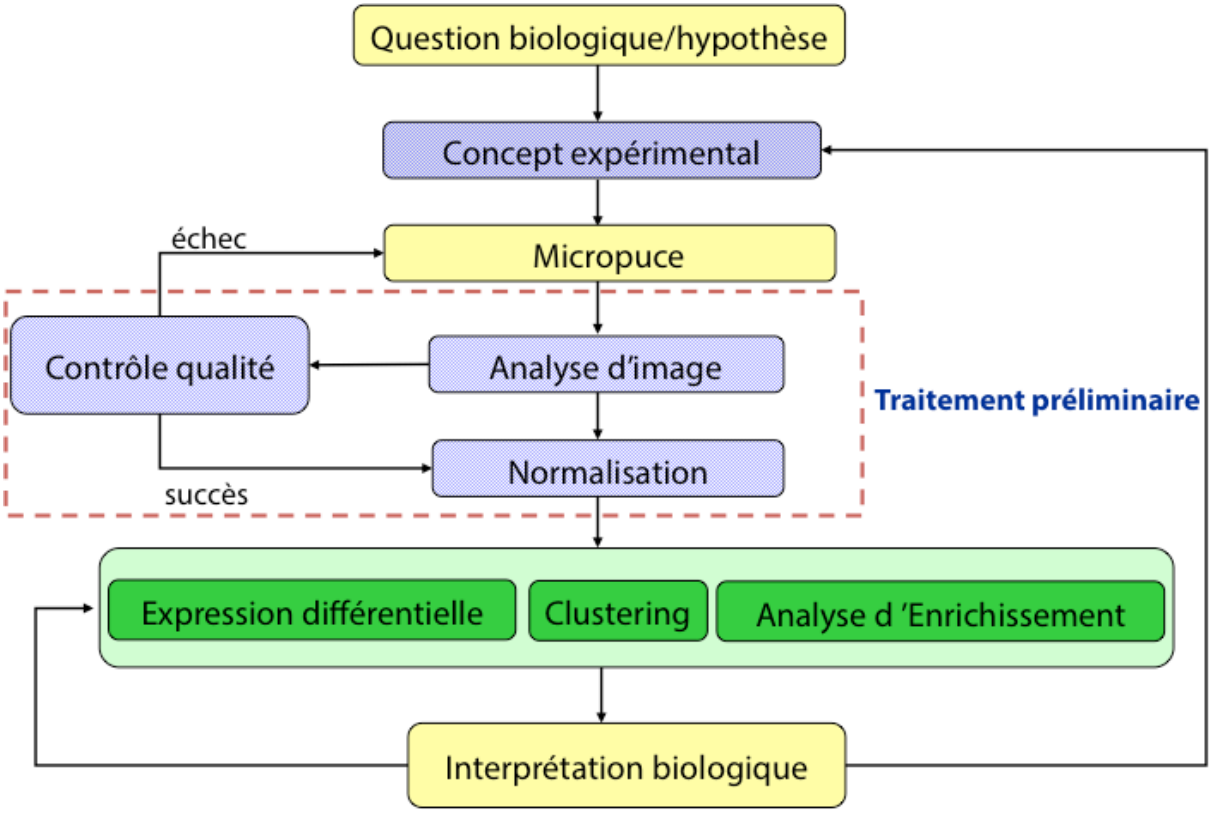

Figure 1.3 Cette figure montre un pipeline pour l'analyse computationnelle des puces d'ADN. Dans la plupart des cas, le but est de déterminer les gènes différentiellement exprimés dans deux conditions différentes pour comprendre les mécanismes moléculaires pathologiques.

1.2.3 RNA-Seq et pharmacogénomique

Je n'ai pas mis en perspective les technologies de séquençage de nouvelles générations de type RNA-Seq puisque les données présentées dans cette thèse sont issues de grandes études transcriptomiques générées sur micropuces. A notre connaissance, les seules données de

RNA-seq disponibles sont ceux de l'étude CCLE (Cancer Cell Line Encyclopedia) pour un sous-ensemble de 935 lignées cellulaires qui ont été publiées via le Cancer Genomics Hub (CGHub), mais sans avoir l'objet d'aucune publication à ce jour. Genentech a aussi récemment publié les données RNA-seq d'un ensemble de 675, mais avec peu de données pharmacogénomiques (Klijn et al. 2015); ces données n'ont pas été utilisées dans la thèse et sont encore en train d'être inspectées dans le laboratoire du Dr. Haibe-Kains. Puisque les études de toxicogénomique et de perturbation transcriptomique tel que TGGATES et L1000, génèrent plusieurs milliers/millions de profils d'expression, il est adéquat d'utiliser la technologie de micropuce en raison de son prix peu élevé par rapport au RNA-seq. RNA-Seq est une technologie relativement nouvelle pour la plupart des chercheurs, et les outils pour l'analyse des données RNA-Seq sont beaucoup plus compliqués que ceux utilisés pour les micropuces. Le décalage entre le développement d'outils d'analyse de données et la rapidité avec laquelle la technologie RNA-seq progresse crée déjà une grande limitation pour les chercheurs surtout au niveau stockage et manipulation des données. Une étude publiée par Wang et al. (Wang et al. 2014) montre les avantages et inconvénients de chaque technique, aussi elle signale que les deux technologies ont une performance similaire vis à vis de la prédiction du mécanisme d'action des produits médicamenteux. Nous préférons traiter chaque technique à part et ne pas appliquer un modèle de mélange, puisque la combinaison des données, issus des ces différentes technologies, est complexe et les méthodes toujours en cours de développement (Thompson et al. 2016).

1.3 Prédiction de la réponse aux agents anticancéreux par analyse de l'expression génique

1.3.1 La pharmacogénomique du cancer

Le développement de la médecine de précision (médecine personnalisée) est devenu une priorité pour la recherche contre le cancer. Cela nécessite l'analyse des données génomiques, à

grande échelle, des individus et des populations afin d'identifier les composantes génétiques qui permettent de caractériser un type de cancer, y compris la probabilité de progression (valeur pronostique) de la maladie et la réponse au traitement. La médecine de précision a profité des avancées dans les technologies génomiques, telles que l'expression des gènes (transcriptomique), le séquençage de nouvelle génération, ainsi que des données cliniques prédictives de l'évolution de la maladie ou de la réponse thérapeutique chez les patients. Ces données sont utilisées pour identifier les gènes et les voies biochimiques dérégulés dans le but de comprendre les facteurs moléculaires qui stimulent la progression tumorale et contribuent à la réponse du patient au traitement. Compte tenu de l'omniprésence de ces ensembles de données, il est maintenant possible d'étudier les sous-types de cancer et d'identifier les aberrations communes et récurrentes dans les cancers. Cette notion a suscité un nouvel intérêt dans le développement et le repositionnement de médicaments anticancéreux pour cibler des aberrations génétiques spécifiques (Goodspeed, Heiser, Gray, & Costello, 2016)

.

1.3.2 Lignées cellulaires cancéreuses

Les lignées cellulaires servent de modèle pour l'étude de la biologie du cancer. Ainsi, les altérations génomiques peuvent aider à comprendre pourquoi certains types de cancer répondent mieux que d'autres aux thérapies anticancéreuses. Par conséquent, un grand nombre de données ont été générées pour corréler les profils génomiques aux réponses pharmacologiques des lignées cellulaires (Figure 1.4). La première étude pionnière, NCI-60, a généré un grand nombre d'essais pharmacologique dans 60 lignées cellulaires cancéreuses (Shoemaker, 2006). Par la suite, les caractéristiques génomiques de ces lignées cellulaires ont été caractérisées et les données NCI-60 ont été compilées dans une ressource publique appelée CellMiner (Shankavaram et al., 2009). Des études menées dans des lignées cellulaires de cancer du sein ont révélé des voies biochimiques et processus biologiques directement touchés par des composés anticancéreux (Heiser et al., 2012). Cependant, la diversité moléculaire du cancer est bien plus étendue que la collection de 60 lignées cellulaires du NCI-60. Pour pallier ce problème, d'autres études pharmacogénomiques a grande échelle telle que le GDSC (Yang et al., 2013), CCLE (Barretina et al., 2012), CTRP (Basu et al., 2013a), ont testé un plus grand

ensemble de lignées cellulaires représentant des dizaines de types de cancer, ce qui a permis une meilleure représentation des types moléculaires observés en clinique. Ces études ont conduit à des progrès majeurs pour la compréhension de la réponse cellulaire aux médicaments et ont fourni les données nécessaires pour développer des algorithmes de prédiction qui visent à correspondre la réponse thérapeutique aux caractéristiques génomiques.

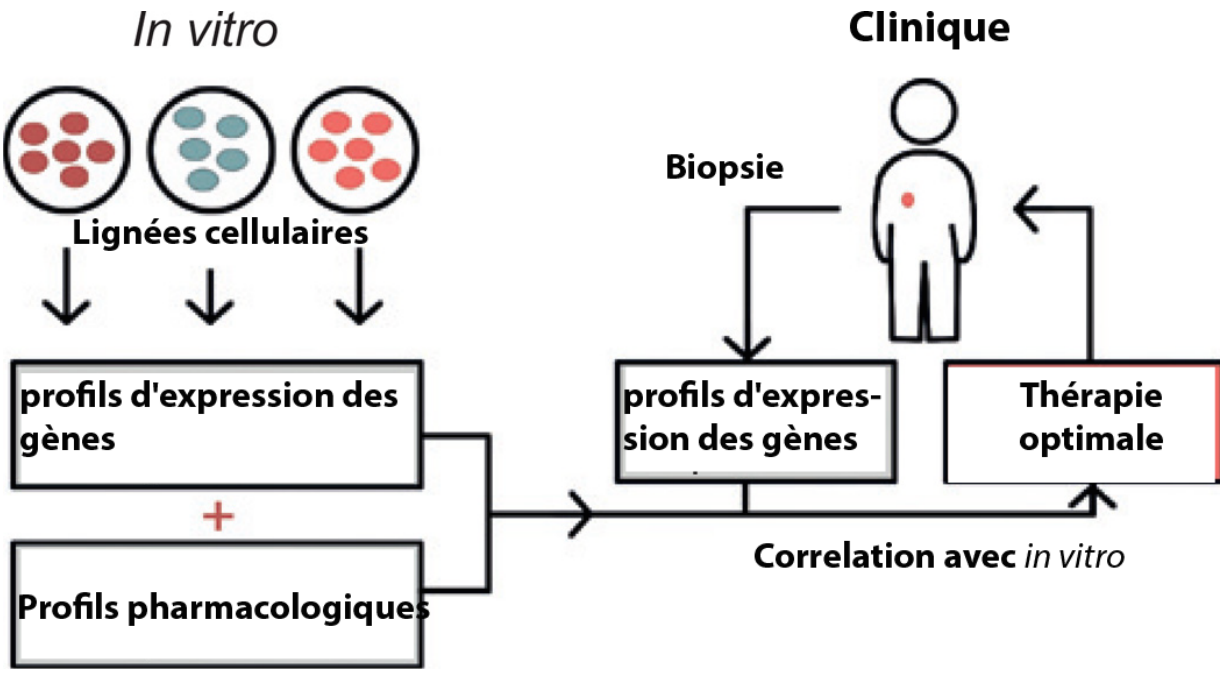

Figure 1.4 cette figure illustre l'utilisation de la génomique/transcriptomique in vitro pour la sélection du traitement clinique. Plusieurs algorithmes de prédiction ont été utilisés pour déterminer les caractéristiques moléculaires prédictives de la réponse pharmacologique des cellules in vitro. Idéalement, les signatures de prédiction dérivées in vitro peuvent être utilisées pour déterminer le meilleur traitement pour la tumeur de l'individu.

1.3.3 Bases de données publiques

NCI-60

NCI-60 regroupe un ensemble de 60 lignées de cellules tumorales humaines. Ces cellules sont parmi les mieux caractérisées (Shoemaker, 2006). Les ensembles de données provenant de ces

lignées cellulaires comprennent des mesures d'expression (ARNm), et une étude de mutations et aberrations génétiques. Les données moléculaires et pharmacologiques (~40,000 médicaments) sont disponibles dans CellMiner (http://discover.nci.nih.gov/cellminer/) (Reinhold et al., 2012) et DTP (http://dtp.cancer.gov/mtargets/mt_index.html), respectivement. Des études ont rapporté des résultats de l'analyse intégrative du nombre de copies de l'ADN, le niveau d'expression des gènes, et la sensibilité des médicaments dans NCI-60 (Bussey et al., 2006; Varma, Pommier, Sunshine, Weinstein, & Reinhold, 2014).

The Cancer Cell Line Encyclopedia

The Cancer Cell Line Encyclopedia (CCLE) est une base de données pharmacogénomique qui compile l'expression génomique, le nombre de copies géniques (CNV) et des données de séquençage de l'ADN dans ~1000 lignées de cellules cancéreuses humaines. Elle comprend aussi les profils pharmacologiques pour 24 médicaments anticancéreux dans ~500 lignées cellulaires (http://www.broadinstitute.org/ccle). L'étude CCLE démontre l'existence de prédicateurs génétiques de sensibilité aux médicaments. Cet ensemble de données est d'une grande importance pour développer des biomarqueurs qui pourraient être utilisés en clinique (MacConaill & Garraway, 2010).

Genomics of Drug Sensitivity in Cancer (GDSC)

GDSC est un projet dédié à la découverte de nouvelles cibles thérapeutiques dans le cancer. (http://www.cancerrxgene.org/). C'est une ressource publique contenant des données d'environ 140 médicaments et ~1000 lignées cellulaires. Les composés étudiés comprennent des thérapies cytotoxiques ainsi que des thérapies ciblées. Comme pour CCLE, la grande collection de lignées cellulaires permet de capter l'hétérogénéité génomique des cancers humains. Les données d'expression et les mutations sont corrélées avec la sensibilité aux médicaments dans des lignées cellulaires afin d'identifier les caractéristiques génétiques qui sont prédictives de la réponse aux médicaments.

The Cancer Therapeutics Response Portal (CTRP)

Initialement, CTRP (http://www.broadinstitute.org/ctrp/) offrait l'accès à des mesures quantitatives de sensibilité pour 354 médicaments et 242 lignées cellulaires (Basu et al.,

2013b). Ce nombre passe à 480 médicaments et 860 lignées cellulaires dans la nouvelle version de CTRP (v2) (Seashore-Ludlow et al., 2015a). Bien que CTRP ne fournit pas des données de profils moléculaires, leurs lignées cellulaires sélectionnées proviennent de la même biobanque que CCLE (Dr Paul Clemons, communication personnelle). Les profils moléculaires de CCLE ont donc été utilisés dans l'étude CTRPv2 pour identifier les médicaments avec un mécanisme d'actions similaire et pour développer de nouveaux biomarqueurs de la réponse aux médicaments (Rees et al., 2016).

1.3.4 Biomarqueurs prédictifs et limitations

Depuis des années, la collection NCI-60 était un point de départ des études prédictives. Le groupe du Dr John Weinstein a utilisé NCI-60 pour explorer la relation entre les marqueurs génétiques moléculaires et la réponse aux médicaments. Par exemple, leurs analyses de variations du nombre de copies des gènes (CNV), expressions des gènes et les données de cytotoxicité ont suggéré que le gène ASNS pourrait être un biomarqueur de la réponse à l'enzyme L-asparaginase (L-ASP) dans le cancer de l'ovaire (Scherf et al., 2000). Une analyse plus poussée en utilisant un ensemble de lignées cellulaires de cancer ovarien a montré une corrélation négative entre le taux de protéine ASNS et la réponse thérapeutique (Lorenzi et al., 2008). En plus, des études fonctionnelles ont indiqué que le knockdown de ASNS sensibilise ces lignées de cellules cancéreuses au traitement à la L-ASP (Lorenzi et al., 2006). Dans une autre étude, l'analyse des mutations dans 24 gènes associés aux cancers connus a contribué à identifier une mutation BRAF, V600E, qui a été associée à la réponse aux phénothiazines (un médicament antipsychotique) dans le mélanome (Ikediobi et al., 2006).

Une étude récente menée sur un grand ensemble de cellules cancéreuses a montré que l'expression de seulement 31 % des cibles de médicaments anticancéreux était corrélée avec la réponse de la cellule cancéreuse (Rees et al. 2016). Il existe donc d'autres déterminants transcriptomiques/génomiques qui affectent la réponse des médicaments dans les cellules cancéreuses. Un exemple cité est celui de l'expression de NQO1, un gène qui potentialise l'effet de la tanespimycin dans les cellules cancéreuse, et qui est un biomarqueur de réponse.

Bien que la tanespimycin cible les chaperonnes de type HSP90; l'expression de HSP90 ne corrèle pas avec la réponse de la cellule cancéreuse.

Les auteurs de l'étude (CCLE) ont corrélé les profils moléculaires de ~1000 lignées cellulaires provenant de 36 types de tissus aux données de réponses aux médicaments en utilisant une technique moderne d'apprentissage machine, nommée Elastic Net (Zou & Hastie, 2005) pour identifier des biomarqueurs génomiques. Ils ont trouvé que l'expression du gène AHR est associée à la réponse d'une lignée cellulaire, présentant un gène mutant NRAS, au traitement par un inhibiteur de la kinase MEK, et l'expression du gène SLFN11 était corrélée avec la sensibilité aux inhibiteurs de la topoisomérase. L'étude GDSC a utilisé la même technique d'apprentissage machine (~650 lignées cellulaires et 130 médicaments). Cette étude a identifié le réarrangement génétique EWS-FLI1 comme associé à l'efficacité d'un inhibiteur de la protéine PARP (olaparib) dans certains sarcomes.

Bien que plusieurs études ont été publiées et ont fait la démonstration du pouvoir prédictif des signatures génomiques pour l'orientation des options thérapeutiques dans différents cancers, peu de biomarqueurs ont pu être validés dans des modèles animaux ou des essais cliniques (Kern, 2012). En conséquence, la reproductibilité des biomarqueurs prédictifs de la réponse aux médicaments anticancéreux est devenue un très grand défi pour la recherche en médecine personnalisée. Une étude récente réalisée par notre équipe a montré que la reproductibilité des biomarqueurs génomique (en utilisant différentes méthodes d'apprentissage machine) est loin d'être parfaite pour 470 lignées cellulaires cancéreuses et 15 médicaments, communes aux deux études CCLE et GDSC (Papillon-Cavanagh et al., 2013a). Les modèles prédictifs (biomarqueurs univariés et multivariés) furent développés sur les lignées dans GDSC et testés sur CCLE pour déterminer la robustesse et la consistance (concordance) de ces biomarqueurs génomiques à prédire le phénotype (sensible vs résistant). Cette analyse a démontré qu'il n'était possible d'identifier des prédicateurs génomiques de réponse que pour une minorité de médicaments (~1/3) alors que pour la majorité des médicaments, les résultats de la validation dans CCLE étaient médiocres (Papillon-Cavanagh et al., 2013b). Ces résultats ont plus tard été confirmés par plusieurs groupes de recherches indépendants (Cortes-Ciriano et al., 2015; S. Dong et al., 2015; Jang, Neto, Guinney, Friend, & Margolin, 2014). La raison de cette

inconsistance était mystérieuse. Ceci nous a motivés à lancer un projet comparatif entre ces deux grandes études. Ce projet déjà publié sera discuté dans le Chapitre 2.

1.4 Prédiction du mécanisme d'action des médicaments par analyse de l'expression génique

Le nombre de nouveaux médicaments approuvés par l'industrie pharmaceutique a considérablement diminué au cours des dernières décennies (Booth & Zemmel, 2004). Il faut approximativement ~15 ans (Dimasi, 2001) et 1 milliard de dollars pour mettre un médicament sur le marché (Adams & Brantner, 2006). Il y a deux principales raisons responsables de ce déclin et coût prohibitif. La première est que les stratégies de développement du médicament qui prévalent au sein des sociétés pharmaceutiques restent conservatrices, et sont généralement orientées sur la découverte d'une nouvelle cible thérapeutique ainsi que l'identification d'un composé thérapeutique qui module l'activité de cette cible. Ceci est suivi par un processus lent, coûteux, et nécessitant une validation expérimentale et clinique.

La deuxième raison est le manque d'évaluation systématique des indications thérapeutiques pour un médicament soit en phase de développement soit après mise sur le marché. Certains des médicaments les plus rentables ont été repositionnés pour de nouvelles indications (Ashburn & Thor, 2004). Les exemples classiques incluent Minoxidil (initialement testé pour l'hypertension; maintenant indiqué pour la perte de cheveux), le Viagra (initialement testé pour l'angine de poitrine, maintenant indiqué pour la dysfonction érectile et l'hypertension artérielle pulmonaire), Avastin (initialement indiqué pour le cancer du côlon et le cancer du poumon; plus tard approuvé pour le cancer du sein métastatique). Il est donc crucial d'identifier les voies moléculaires ciblées et le mode d'action (MoA) pour le développement de nouveaux médicaments, et pour de nouvelles applications cliniques de médicaments déjà disponibles (Ambesi-Impiombato & Bernardo, 2006; Terstappen, Schlüpen, Raggiaschi, & Gaviraghi, 2007). Les approches impliquant la biologie des systèmes (approche

computationnelle) sont plus flexibles pour étudier la complexité de l'activité des médicaments dans le contexte cellulaire (Berger & Iyengar, 2009; Mani et al., 2008).

1.4.1 Mécanisme d'action de médicaments

Similarité chimique et MoA

Les propriétés chimiques d'un composé médicamenteux sont évidemment associées à son utilisation efficace (pharmacocinétique, absorption…). Il est donc possible d'explorer les possibilités de repositionnement de médicaments en fonction des caractéristiques chimiques communes. La base rationnelle pour cette approche est présente dans la relation structure — activité biochimique (QSAR). Bien que les structures chimiques similaires ne se comportent pas toujours de la même façon dans les systèmes biologiques, les degrés de similarité existants peuvent être exploités en utilisant des approches computationnelles pour comprendre le mécanisme d'action et le repositionnement des médicaments. L'approche computationnelle consiste à regrouper (« clustering ») les médicaments sur base d'un ensemble de caractéristiques chimiques communes, puis de relier les médicaments les uns aux autres par des réseaux de similarités (Eckert & Bajorath, 2007). Ainsi, de nouvelles indications cliniques peuvent être inférées par association des caractéristiques biologiques, telles que des cibles moléculaires connues et enrichies dans le réseau de médicaments.

Keiser et al. ont implémenté une approche qui intègre à la fois la similarité entre les composés chimiques, ainsi que les interactions médicaments-protéines (cible) bien établies. Dans cette approche, la cible du médicament est représentée par l'ensemble des structures chimiques qui sont connues pour s'associer à celle-ci. Pour évaluer une nouvelle association possible entre un médicament établi et une hors cible (« off-target »), un score a été obtenu en calculant la similarité entre la structure du composé en question et chaque membre de l'ensemble des médicaments qui s'associe à la cible. Le coefficient de Tanimoto étant la mesure qui représente la similarité structurelle entre deux composés chimiques (Figure 1.5). Plusieurs des interactions hors cibles prédites par cette méthode ont été confirmées expérimentalement (Keiser et al., 2009).

Les limites de l'approche utilisant la similarité chimique pour déterminer le mode d'action d'un médicament provenaient du fait que de nombreuses structures chimiques peuvent contenir des erreurs. En outre, de nombreux effets physiologiques ne peuvent être prédits par les seules propriétés chimiques, parce que les médicaments subissent des transformations métaboliques et pharmacocinétiques complexes (Fourches, Muratov, & Tropsha, 2010).

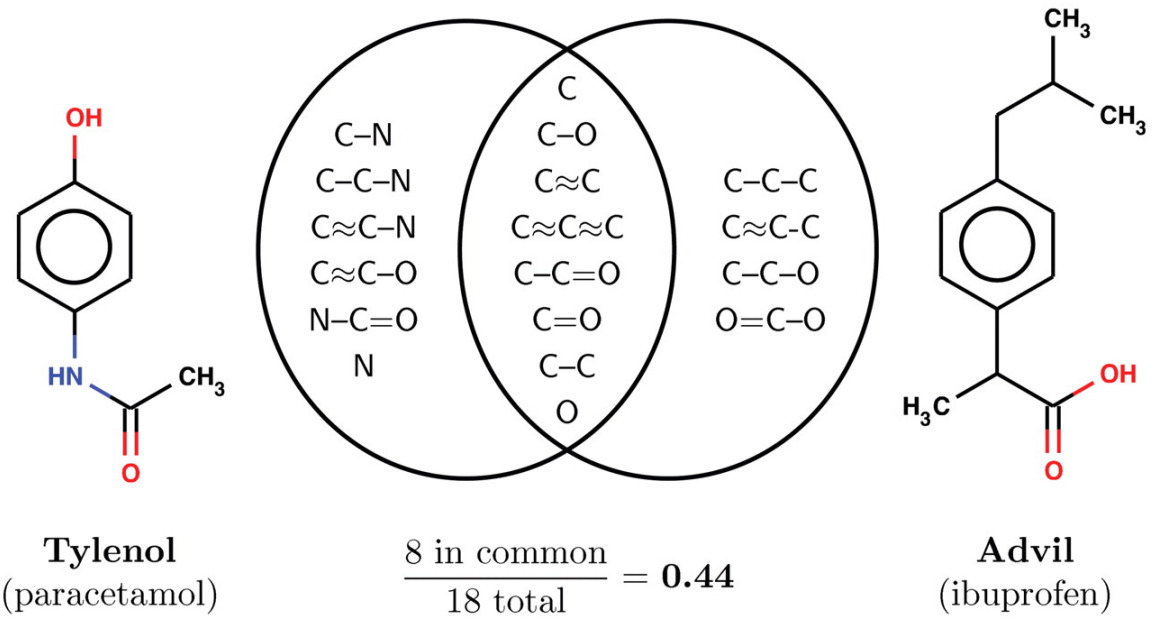

Figure 1.5 cette figure illustre comment calculer l'indice de tanimoto qui est une mesure de similarité entre deux structures chimiques.

Similarité transcriptomique/activité biologique

Lorsqu'un composé, pharmacologiquement actif, est exposé à un système biologique, il en résulte une perturbation moléculaire (Figure 1.6A). Bien que le mécanisme d'action de plusieurs médicaments n'est pas clairement établi, les techniques de mesure moléculaire à haut débit, telle que les micropuces d'ADN, peuvent être utilisées pour mesurer et caractériser l'impact d'un composé sur l'ensemble des gènes dans un système biologique. De cette façon, il devient possible de construire une « signature » de l'activité moléculaire du médicament. Ces signatures de l'activité moléculaire peuvent renseigner sur le mécanisme d'action, même dans les cas où la cible primaire est inconnue.

L'une des approches les plus systématiques et les plus connues est le projet Connectivity Map (CMAP) (Lamb et al., 2006) (voir section suivante). Les stratégies pour établir le mécanisme d'action de médicaments ne sont pas limitées à l'analyse du transcriptome perturbé par le médicament, et peuvent inclure d'autres types de données tels que des essais de criblage de produits chimiques ou des analyses d'essais pharmacologiques à haut débit (Figure 1.6B) et technique d'invalidation génique « knock-out ».

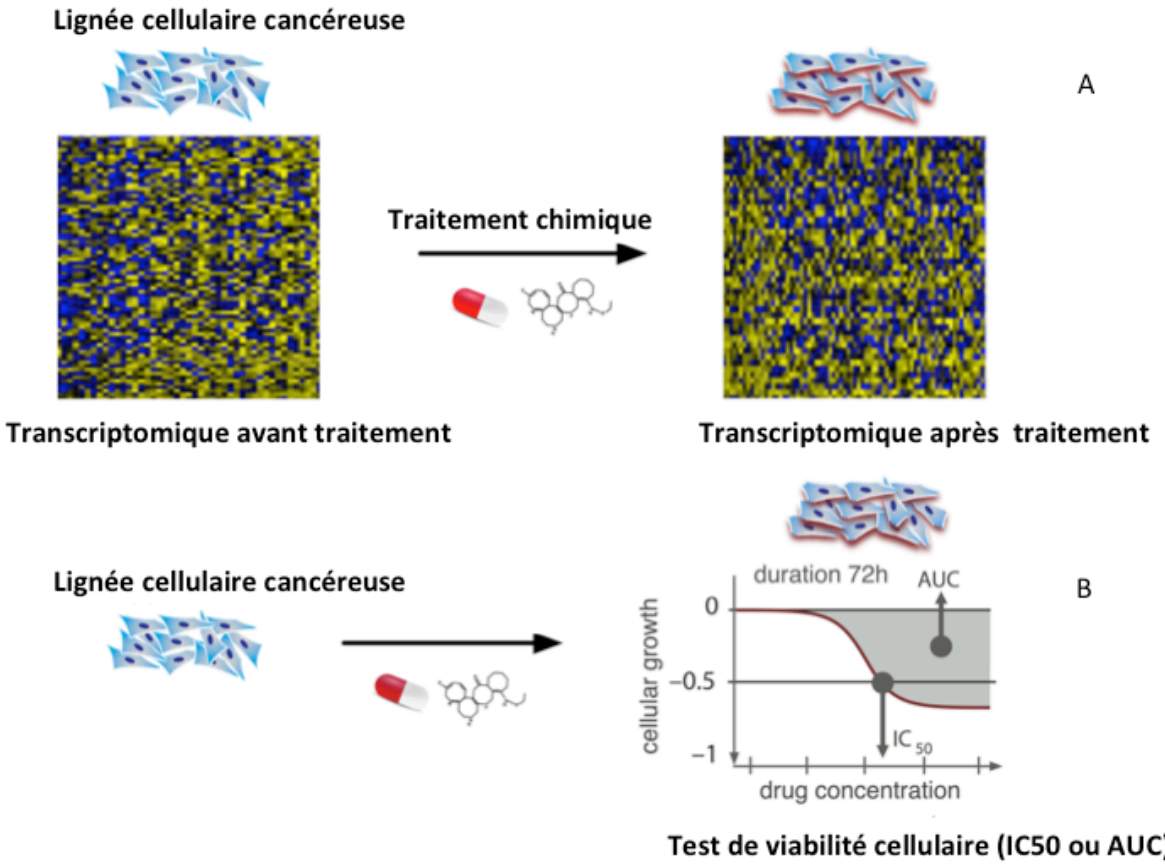

Figure 1.6 cette figure représente le concept de CMAP (A). CMAP a mesuré l'expression des gènes perturbés par un médicament donné dans des lignées de cellules cancéreuses. (B) Représente les concepts NCI60 et CTRPv2, qui ont mesuré l'inhibition de la prolifération des cellules cancéreuses traitées par différentes concentrations de médicaments.

1.4.2 Concept de la « Connectivity Map »

Connectivity Map (CMAP)

La Connectivity Map (http://www.broadinstitute.org/cmap/) est une grande base de données publique contenant des profils d'expression du génome entier de cellules cancéreuses humaines, traitées avec ~1,300 médicaments à des concentrations différentes (Lamb et al., 2006). Les données sont organisées en lots (« batch ») constitués de plusieurs expériences de puces à ADN correspondant aux lignées cellulaires traitées ou non par le médicament (traitement vs contrôle), pour un nombre total de 6,100 « instances ». Une instance se compose d'un traitement et du contrôle correspondant pour un médicament donné. Chaque instance a un certain nombre d'attributs comprenant un identificateur unique, le lot dans lequel il a été produit, le nom CMAP du traitement, la source, la concentration, la lignée cellulaire CMAP utilisée. L'expression différentielle des gènes est mesurée pour un groupe traité par rapport au contrôle pour chaque médicament dans un lot donné.

L'objectif du projet CMAP est de générer un modèle détaillé qui associe les gènes perturbés aux mécanismes de médicaments et applications biomédicales potentielles. Afin d'identifier les gènes et voies biologiques significativement perturbés par le traitement de chaque médicament, l'équipe CMAP a développé un algorithme qui permet d'associer une signature d'intérêt (gènes différentiellement exprimés entre tissus tumoraux et normaux par exemple) aux profils déposés dans CMAP.

LINCS (The Library of Integrated Network-based Cellular Signatures)

Le projet LINCS est un programme géré par le National Institute of Health (NIH) aux États-Unis d'Amérique. Les fonds pour ce projet ont permis la production d'environ un million de profils d'expression des gènes en utilisant la technologie L1000. La technologie de profilage d'expression génique peu coûteuse, nommée L1000, a été développée pour la nouvelle version de CMAP par le même groupe au Broad Institute. Cette plateforme mesure environ 1000 gènes dans chaque expérience. Le reste du transcriptome, environ 20,000 gènes, est estimé par un modèle statistique à partir de milliers de jeux de données d'expression génique dans Gene Expression Omnibus (GEO) (Barrett & Edgar, 2006). Le raisonnement derrière ce modèle est

que les 1000 gènes sélectionnés permettent de raisonnablement inférer le reste du transcriptome qui n'est pas directement mesuré expérimentalement.

Le technique de la plate-forme L1000 commence par l'amplification d'ARNm à partir de lysats cellulaires en utilisant une amplification médiée par ligature (Peck et al., 2006). Les sondes contiennent deux parties, une en amont et une autre en aval de la séquence pour des gènes spécifiques. Les sondes sont ensuite hybridées à des séquences constituées de l'acide désoxyribonucléique complémentaire (ADNc), créé à partir des ARNm isolés, puis ligaturées par une ligase Taq. Les sondes sont ensuite amplifiées par réaction en chaîne (PCR) et hybridées à des billes Luminex. Les billes sont ensuite détectées et quantifiées en utilisant la technologie Luminex FLEXMAP 3D® (Roth & Wayne, 2010). Cette technologie utilise des faisceaux laser pour détecter l'identité des billes. Outre les ~1000 gènes sélectionnés, 80 autres transcrits invariants sont mesurés pour normaliser les mesures d'expression des gènes dans chaque plaque. Le nombre réduit de variables maintient un coût réduit pour estimer l'ensemble du transcriptome. Cela rend la technologie appropriée pour remplir l'objectif du projet LINCS, à savoir d'analyser les changements transcriptomiques qui se produisent lorsque les cellules sont exposées à de nombreux agents perturbateurs.

Jusqu'à présent, les données qui ont été collectées par la technologie L1000 comprennent ~ 22,412 perturbations uniques appliquées à 56 contextes cellulaires, y compris des lignées de cellules de tissus sains et des lignées de cellules cancéreuses humaines. ~16,425 des perturbations sont des composés chimiques y compris des médicaments, des ligands et d'autres petites molécules appliquées à doses et durée d'exposition différentes. ~5800 des perturbations sont d'ordre génétique, c'est-à-dire la réduction ou surexpression de gènes uniques. Les expériences d'invalidation génique « knockdown » utilisent la technologie de l'ARN interférant (RNAi) pour cibler les gènes. Le nombre total de profils d'expression génique mesurée à ce jour est d'environ 1.8 million (Duan et al., 2014). Comme toutes les expériences ont été réalisées sur des cellules humaines en utilisant des ensembles bien définis de perturbations, en utilisant la même plate-forme, il est possible de développer des méthodes computationnelles qui seraient en mesure d'analyser et d'interpréter ce grand ensemble de données dans le but de mieux comprendre le mode d'action de milliers de médicaments.

1.4.3 Études pertinentes et limitations

La première étude pionnière qui a démontré l'utilité des profils d'expression des gènes (perturbés par des centaines de médicaments dans CMAP) pour comprendre le mécanisme d'action des médicaments est paru en 2010 (Iorio et al., 2010). Les auteurs de cette étude ont développé une procédure générale pour prédire le MoA de nouveaux composés, et trouver des nouvelles applications biomédicales, connues sous le terme repositionnement de médicament ou « drug repurposing ». L'originalité de cette étude provient du fait que les auteurs ont identifié une réponse transcriptionnelle « consensus » d'un médicament à travers les différentes lignées cellulaires de la base de données CMAP. Ils ont ensuite automatiquement extrait une signature transcriptomique (les gènes les plus sur — ou sous-exprimés) pour chaque médicament et ont calculé la similarité entre les médicaments en se basant sur cette signature transcriptomique. Ils ont ensuite analysé le réseau de médicaments résultant afin d'identifier les communautés présentant un même MoA et de déterminer les voies biologiques perturbées par ces médicaments. Ils ont démontré que leur approche regroupait également des médicaments interagissant avec des membres distincts de la même voie de signalisation biologique. Une fois établi, un tel réseau peut être utilisé pour déduire le MoA ainsi que les voies ciblées par les composés anticancéreux encore sous investigation.

Une autre étude parue en 2013 (Napolitano et al., 2013) a suggéré qu'il était insuffisant de considérer l'expression génique induite par les médicaments comme seule source d'information pour prédire le MoA compte tenu des limitations liés à l'analyse des données transcriptomiques. Les auteurs ont proposé une nouvelle approche computationnelle basée sur des algorithmes d'apprentissage automatique. Dans cette analyse ils ont intégré plusieurs couches d'informations concernant les médicaments dans CMAP : la similarité de la structure chimique, la similarité entre deux médicaments si leurs cibles correspondantes sont proches au sein du réseau d'interactions protéine-protéine, et la corrélation de l'expression des gènes. Ils ont démontré que l'intégration de plusieurs types d'informations était meilleure (statistiquement significative) que si l'on considère l'expression des gènes seule pour la prédiction des indications thérapeutiques des médicaments approuvées.

Récemment, les membres du laboratoire du Dr Andrea Califano ont mis au point une nouvelle approche appelée « DemanD » pour caractériser les MoA d'un médicament donné (Woo et al., 2015). Le procédé implique la création d'un modèle bioinformatique du réseau d'interactions géniques spécifiques dans une cellule cancéreuse (tumeur vs normal). Les expériences sont alors réalisées pour analyser les changements dans l'expression de gènes dans ces cellules cancéreuses exposées à un médicament d'intérêt. L'algorithme combine alors les données provenant du réseau de gènes du modèle pathologique avec des données d'expressions différentielles générées après traitement par le médicament dans les mêmes cellules. Ce modèle permet l'identification de différents régulateurs impliqués dans le MoA. Malgré les résultats prometteurs de ce modèle, une des limites importantes est que son application est peu pratique pour un très grand nombre de médicaments et lignées cellulaires, car la performance de ce modèle dépend largement de la qualité du réseau de régulation spécifique à un certain contexte cellulaire, ainsi que le profil d'expression après traitement dans une seule lignée cellulaire.

Finalement, une étude très récente utilisant le nouveau jeu de données LINCS L1000 a démontré qu'il existe une connexion entre structure chimique, activité biologique et expression de gènes induits par un médicament donnée (Chen et al., 2015). Cette étude menée sur 11,000 médicaments suggère qu'il serait intéressant d'intégrer toutes les informations de base concernant les médicaments pour identifier leur MoA. L'intégration de ces données sera le sujet du Chapitre 3 de la thèse.

1.5 Mécanismes d'hépatotoxicité et analyse de l'expression génique

Au cours de la dernière décennie, l'industrie pharmaceutique a souffert de taux d'attrition élevés pour mettre sur le marché une nouveau médicament ou molécule thérapeutique, notamment en raison des effets secondaires néfastes au cours des phases avancées des essais cliniques. Ces résultats décevants sont surtout liés à l'efficacité du médicament (spécificité de la cible), les propriétés d'absorption, de distribution, de métabolisme et d'excrétion (ADME). Dans 30-40 % des cas ceci est également lié à une toxicité qui se manifeste chez l'humain, en

particulier pour le foie et le cœur, en dépit du fait que les tests sur les animaux n'ont pas rapporté d'effets toxiques majeurs (Kola & Landis, 2004). La toxicité inattendue/idiopathique chez l'homme peut même se présenter après la mise sur le marché. Chaque année, environ deux millions de patients aux États-Unis subissent un effet indésirable grave lors de l'utilisation des médicaments commercialisés, entraînant approximativement 100,000 décès, ce qui représente la quatrième cause principale de décès (Giacomini et al., 2007). Cela suggère qu'un modèle animal standard pour l'évaluation de la sécurité chimique manque clairement de sensibilité. L'échec dans les dernières phases de développement du médicament est évidemment au détriment des patients, mais aussi, compte tenu des coûts extrêmes de développement de nouveaux médicaments, implique d'énormes pertes économiques (Paul et al., 2010).

En même temps, d'autres exemples montrent que les modèles animaux pour tester la toxicité chronique (doses répétées) peuvent également représenter une fausse piste pour l'évaluation des risques pour la santé humaine : comme rapporte le US Physicians Desk aux États-Unis, ainsi sur 241 agents pharmaceutiques utilisés pour les traitements chroniques, 101 agents ont été démontrés cancérigènes sur les rongeurs. Cependant, des données épidémiologiques chez les patients traités de manière chronique telles que revues par l'Agence internationale pour la recherche sur le cancer ont identifié seulement 19 de ces produits pharmaceutiques (8 %), principalement destinés au traitement anticancéreux ou hormonal, comme cancérigènes chez l'humain. Ce manque apparent de spécificité et de sensibilité de l'essai toxicologique chez les rongeurs est souligné par un rapport indiquant que seulement 43 % des effets toxiques induits par des produits pharmaceutiques chez les humains ont été correctement prédits par des tests chez les rongeurs (Hartung, 2009). Ces exemples montrent que l'industrie pharmaceutique est en recherche de nouvelles méthodologies et meilleures alternatives permettant de prédire la toxicité des médicaments chez les humains.

1.5.1 La toxicogénomique

Le domaine de la toxicologie a pu profiter de l'avènement des technologies de la génomique ce qui créa une nouvelle discipline, la toxicogénomique (Nuwaysir, Bittner, Trent, Barrett, &

Afshari, 1999), dans l'espoir de fournir des alternatives aux modèles animaux actuels pour l'évaluation de la toxicité des médicaments. Ceci est, par exemple, exprimé dans le règlement de l'Union européenne 2006 sur l'enregistrement, l'évaluation et l'autorisation des produits chimiques (REACH), qui traitent de la production et l'utilisation des substances chimiques, ainsi que leurs impacts potentiels sur la santé humaine et l'environnement. La législation REACH énonce :

> « *The Commission, Member States, industry and other stakeholders should continue to contribute to the promotion of alternative test methods on an international and national level including computer supported methodologies, in vitro methodologies, as appropriate, those based on toxicogenomics, and other relevant methodologies.* »

L'approche principale pour le développement des essais prédictifs basés sur la toxicogénomique, en particulier aux fins d'identification des effets toxiques néfastes chez l'humain, implique que les données génomiques soient dérivées de l'exposition des échantillons biologiques à des substances toxiques connues. L'essai biologique peut se référer à des modèles animaux, mais pour les essais *in vitro*, les modèles cellulaires humains représentent un outil de choix. Par critère de toxicité, les composés prototypiques sont dérivés de bases de données toxicologiques disponibles telles que celle du US National Toxicology programme (http://ntp.niehs.nih.gov/). Cependant, il n'y pas encore de consensus en ce qui concerne l'essai biologique et les méthodes computationnelles pour générer un ensemble prédictif de gènes permettant de classifier un composé selon un phénotype toxique particulier. Afin de tester la validité d'un tel modèle prédictif, une validation peut être effectuée sur un ensemble de composés toxiques modèles. Il est évident que la sélection du plus grand nombre de composés prototypiques appartenant à différentes classes chimiques, à partir des bases de données disponibles, est cruciale pour améliorer la précision du modèle prédictif (Vinken et al., 2008).

De manière similaire à la Connectivity Map, les changements au niveau de l'expression des gènes, induits par une substance toxique présumée, sont comparés à une série de changements induits par d'autres composés toxiques. Si les caractéristiques correspondent, un certain mode d'action toxique peut être attribué à l'agent inconnu, identifiant ainsi un risque potentiel pour la santé humaine. À ce jour, l'analyse de l'expression des gènes au niveau de l'ensemble du

génome, en appliquant la technologie des micropuces à ADN, est la technique dominante en toxicogénomique.

1.5.2 Le foie, site majeur de détoxification

Le foie présente des propriétés biochimiques uniques et se positionne comme le principal organe impliqué dans le métabolisme de nombreux xénobiotiques (produits chimiques) (Fasinu, Bouic, & Rosenkranz, 2012). Comme conséquence, il est le principal organe qui est affecté par des doses nocives de xénobiotiques, ce qui entraîne une toxicité hépatique. De fait, le foie humain est le candidat de choix pour un modèle d'étude de la toxicité humaine en réponse à l'exposition des xénobiotiques. Les hépatocytes constituent le type de cellules majoritaire du foie; environ 60 % des cellules sont des hépatocytes. Ces cellules contribuent à plusieurs fonctions du foie. L'autre population cellulaire hépatique est représentée par des cellules moins volumineuses. Ces cellules comprennent : les cellules endothéliales tapissant les vaisseaux sanguins et voies biliaires, les cellules de Kupffer, les cellules immunitaires, les fibroblastes et les cellules souches (de Graaf et al., 2010). Plusieurs modèles cellulaires liés au foie humain ont été développés pour l'étude du métabolisme et de la toxicité des médicaments.

Lignées cellulaires hépatiques immortalisées

Plusieurs modèles *in vitro* à partir de ces cellules ont été mis au point, deux des plus populaires étant les cellules HepG2 et HepaRG. Alors que la lignée cellulaire HepG2 a été disponible depuis le début des années 1980 (Knowles, Howe, & Aden, 1980; Morris, Aden, Knowles, & Colten, 1982), la lignée cellulaire HepaRG a été ajoutée plus récemment aux modèles *in vitro* (Aninat et al., 2006; Cerec et al., 2007). Les cellules immortalisées présentent certains avantages majeurs tel que la capacité à se diviser rapidement, ce qui permet une meilleure accessibilité à ces cellules à des coûts financiers réduits. Cela a conduit à l'adoption de ces cellules à l'échelle mondiale, ce qui entraîna une grande disponibilité des données dérivées de ces modèles *in vitro*, en particulier pour HepG2. Les données disponibles sur l'exposition de ces cellules à des xénobiotiques suggèrent qu'ils répondent aux exigences des études de toxicité (Jennen et al., 2010; van Delft et al., 2004). Cependant, il existe aussi

des inconvénients a l'utilisation de ces modèles cellulaires *in vitro*. L'un d'entre eux étant que, par rapport au foie, ces cellules présentent des niveaux inférieurs de plusieurs des enzymes qui métabolisent des xénobiotiques (enzymes de phase I et enzymes de biotransformation de phase II) et des transporteurs cellulaires (Olsavsky et al., 2007; Westerink & Schoonen, 2007). Il est à noter cependant que sur les modèles cellulaires mentionnés ci-dessus, les cellules HepaRG en général sont soupçonnés d'avoir des activités enzymatiques de phase I qui sont comparables à celles du foie humain (Kanebratt & Andersson, 2008; Lambert, Spire, Renaud, Claude, & Guillouzo, 2009). L'expression aberrante des enzymes métabolisant les xénobiotiques peut être une conséquence du contexte tumoral de ces modèles *in vitro*.

Hépatocytes humains primaires

Les hépatocytes humains primaires (PHH) sont des cellules hépatiques dérivées directement à partir du foie humain, qui sont largement capables de conserver leur fonctionnalité (Hewitt et al., 2007; Olsavsky et al., 2007). Ils peuvent être mis en culture immédiatement après l'isolement du foie, ou ils peuvent être congelés et conservés dans de l'azote liquide pour être mis en culture ultérieurement. Pour une performance optimale, les PHH doivent être mis en culture en utilisant une configuration en sandwich. Dans cette configuration, ces cellules sont munies d'un gel-matrice (« matrigel »). Ainsi, les cellules sont entourées d'un environnement en 3D mimant la situation *in vivo* dans le foie. En conséquence, ceux-ci semblent conserver la plupart de leurs fonctions hépatiques spécifiques y compris l'expression génétique et l'activité enzymatique, et la capacité à former des canalicules biliaires (LeCluyse, Witek, Andersen, & Powers, 2012; Swift*, Pfeifer*, & Brouwer, 2010). Les cultures primaires d'hépatocytes humains peuvent provenir d'un ou plusieurs donneurs et peuvent être regroupées en une seule culture. Ainsi, les hépatocytes primaires humains permettent l'étude des différences interindividuelles en réponse à l'exposition des xénobiotiques. Cependant, il existe également certaines limites à l'utilisation de ce modèle. Tout d'abord, la durée de vie des hépatocytes humains primaires est limitée à quelques semaines. En plus, ce sont des cellules complètement différenciées et ne se divisent donc pas, ce qui est une limitation importante (Michalopoulos, Bowen, Nussler, Becich, & Howard, 1993). Enfin, comme les modèles cellulaires immortalisés, les hépatocytes humains primaires ne représentent qu'un seul type cellulaire alors qu'un modèle contenant plusieurs types cellulaires représentatifs du foie serait optimal.

L'utilisation de cultures primaires d'hépatocytes humains est préférentielle pour prédire la toxicité *in vivo* chez l'Humain, mais elle est entravée par la disponibilité des échantillons cliniques et la variabilité génétique interindividuelle (LeCluyse 2001). Les hépatocytes humains sont choisis comme modèle expérimental de référence en raison des propriétés suivantes : 1. Ils retiennent les mêmes compétences métaboliques (enzymes métaboliques et les cofacteurs) que le foie chez l'Homme. 2. Les hépatocytes humains retiennent tous les aspects physiologiques pertinents pour étudier les cibles toxicologiques Les hépatocytes sont le principal type de cellules endommagées lors des effets indésirables suite à l'administration de médicaments (lésion hépatique attribuée aux xénobiotiques). Les lésions hépatocellulaires étant l'événement initiateur qui, à la suite de multiples facteurs de risque, conduirait à une cascade de réactions toxicologiques, qui culminent avec l'insuffisance hépatique sévère (Amacher 2012). Les hépatocytes primaires ont aussi permis de reproduire certains types de toxicité, comme la formation de métabolites réactifs, la phospholipidose, ou des altérations de type cholestatique. Les paramètres choisis pour l'évaluation toxicologique incluent : la formation d'espèces oxygénées réactives (ROS), l'épuisement de l'ATP, l'activation des caspases et l'appauvrissement du glutathion permettent respectivement de quantifier le stress oxydatif, les dommages hépatocellulaires, l'apoptose et la formation de métabolites réactifs (Li 2002). Ces événements sont généralement considérés comme des mécanismes toxicologiques clés associés aux lésions hépatiques. Une préoccupation souvent soulevée par l'évaluation des lésions hépatiques attribuées aux médicaments utilisant les systèmes *in vitro* est que la manifestation clinique impliquerait des interactions complexes *in vivo* entre les hépatocytes et autres types cellulaires, ce qui est difficile à modéliser adéquatement en utilisant des systèmes *in vitro*, tels que le syndrome de la voie biliaire ainsi que la toxicologie liée au système immunitaire (Chen et al. 2013). Une étude récente a démontré que le ratio ROS/ATP identifié, avec une sensitivité et spécificité raisonnable, les médicaments hépatotoxiques qui induisent directement des lésions hépatocellulaires ce qui pourrait ainsi être facilement détecté dans un système d'hépatocytes humains (Zhang et al. 2016).

1.5.3 Bases de données toxicogénomiques

Compte tenu du coût important de la toxicogénomique à grande échelle, la plupart des études ont généré des informations limitées pour l'application d'approches bioinformatiques sophistiquées, nécessitant un grand nombre d'expériences pour la validation statistique. En 2011, deux grandes bases de données de toxicogénomique furent rendues publiquement accessibles : le projet toxicogénomique japonais (TG-GATES, aka Toxicogenomics Project-Genomics Assisted Toxicity Evaluation system) et DrugMatrix. TG-GATES fut réalisée par l'Institut national japonais des sciences de la santé, l'institut national de l'innovation biomédicale, et 15 entreprises pharmaceutiques (http://toxico.nibio.go.jp/open-tggates/search.html) (Uehara et al., 2010). DrugMatrix fut généré par Iconix Pharmaceuticals (Ganter et al., 2005) et fut rendu public par l'Institut national des sciences de la santé de l'environnement des États-Unis (https://ntp.niehs.nih.gov/drugmatrix/index.html).

Ces deux bases de données mettent l'accent sur les médicaments commercialisés et les données d'expression des gènes à partir de cellules hépatiques. Certains médicaments ont été testés à la fois par TG-GATES et DrugMatrix en doses et durées multiples, *in vitro* et *in vivo*. Les plus grands avantages de ces deux bases de données par rapport aux autres bases de données telles que GEO (Barrett & Edgar, 2006), ArrayExpress (Sarkans et al., 2005), CEBS (M. Waters et al., 2003), et autres (Burgoon, Boutros, Dere, & Zacharewski, 2006; Hayes et al., 2005) sont les suivants :.

- La conception expérimentale uniforme rend l'étude comparative plus simple et pertinente entre les traitements chimiques
- Le grand nombre de médicaments commercialisés et testés fournit une occasion sans précédent pour l'évaluation globale des modèles précliniques basée sur les puces d'ADN pour prédire la toxicité chez l'humain.
- Tester à la fois *in vitro* et *in vivo* pour le même ensemble de produits chimiques permet de déterminer la similarité et la différence entre les deux systèmes en toxicologie prédictive.

TG-GATES

TG-GATES a testé 170 composés chimiques, principalement des médicaments. Leur principal organe cible est le foie, mais aussi des échantillons de rein sont disponibles. Les données de ~20,000 micropuces d'ADN furent générées à la fois *in vitro* et *in vivo* (Table 1.1). Les expériences *in vivo* utilisent l'espèce de rats Sprague Dawley mâles avec deux modèles expérimentaux différents, dose unique et étude à doses répétées. Pour les deux modèles, la dose maximale tolérée (DMT) fut déterminée après 1 semaine de traitement avec chaque composé. Basé sur le poids corporel, le poids des organes, et l'anatomopathologie), la DMT fut définie comme la dose la plus élevée dans l'étude à dose répétée. Dans l'étude à dose unique, presque tous les composés furent administrés à la même dose pour faciliter la comparaison de l'expression génique entre les études à doses simples et répétées. Plus précisément, dans l'étude à dose unique, les rats furent traités selon trois niveaux de dose (faible, moyenne et élevée), à chaque groupe traité correspond un groupe contrôle. Les rats furent sacrifiés à 3, 6, 9 ou 24 heures. Dans l'étude à doses répétées, les rats furent également traités pour trois niveaux de dose (faible, moyen et élevé), mais sacrifiés 24 heures après la dernière dose répétée pour 3, 7, 14 et 28 jours. Pour chaque groupe (dose-période de traitement), les données d'expression génique furent initialement analysées en tenant compte de trois animaux par groupe. D'autres données obtenues comprennent l'examen histologique, tests enzymatiques, l'hématologie, le poids corporel, le poids des organes, et des symptômes généraux.

En plus des données in vivo, TG-GATES contient deux types d'études *in vitro*, des hépatocytes primaires de rats Sprague-Dawley mâles et des donneurs humains. Ils ont été traités aussi avec trois niveaux de dose (faible, moyenne et élevé), et soumis à l'analyse d'expression génique à 2, 8 et 24 heures après le traitement.

Espèce	Rat male Sprague Dawley			Donneur Humain
Type d'etude	In vivo	In vivo	In vitro	In vitro
Dose	Dose unique	Dose répétée	Dose unique	Dose unique
Niveau de dosage	Contrôle, faible, moyenne, élevée	Contrôle, faible, moyenne, élevée	Contrôle, faible, moyenne, élevée	Contrôle, faible, moyenne, élevée
Collection des échantillons après traitement chimique	3, 6, 9, and 24 h	3, 7, 14, and 28 days	2, 8, and 24 h	2, 8, and 24 h
Réplicats biologiques	Triplicata	Triplicata	Duplicata	Duplicata
Plateforme de Micropuce	Affymetrix RG230-2.0 array	Affymetrix RG230-2.0 array	Affymetrix RG230-2.0 array	Affymetrix human U133 plus 2.0 array
Nombre de composés chimiques	131	131	131	119
Autres tissus	rein		Non	

Tableau 1.1 Ce tableau représente un aperçu de l'organisation de la base de données toxicogénomique TGGATES. Le présent schéma est un exemple des données analysées dans le chapitre 4 de la thèse.

1.5.4 Études et limitations

Les données d'expression génique qui sont générées dans une étude de toxicogénomique sont très volumineuses et complexes (voir section précédente, exemple TG-GATES). Par exemple, une expérience *in vitro* en trois exemplaires biologiques (triplicata), trois groupes de dose et deux durées d'exposition, 3 contrôles (x 2 durées d'exposition) =6 + 3 échantillons traités (x 3 doses x 2 durées d'exposition) = 18 exige en total (18 + 6) 24 micropuces d'ADN en total (2400 pour 100 composés juste pour les expériences *in vitro*) et chaque puce représente l'expression de ~20,000 gènes (transcrits ou mRNA). La collecte minutieuse, la gestion et l'intégration de ces données sont dès lors essentielles pour l'interprétation des résultats toxicologiques (M. D. Waters & Fostel, 2004).

Le défi des études toxicogénomiques est de transformer les données brutes relativement bruitées, pouvant masquer l'interprétation, en des conclusions biologiques pertinentes (Heijne, Stierum, Leeman, & van Ommen, 2005; Stierum, Heijne, Kienhuis, van Ommen, & Groten,

2005). Les méthodes les plus utilisées appliquent une normalisation des données brutes produites par les micropuces d'ADN. Après la normalisation, il sera possible de déterminer les gènes différentiellement exprimés dans le groupe traité versus contrôle non traité (voir Chapitre 1 analyse des puces d'ADN).

La relation entre les gènes dans un contexte biologique peut efficacement augmenter les informations biologiques qui peuvent être récupérées à partir d'une seule expérience de toxicogénomique (Currie, Orphanides, & Moggs, 2005). Plusieurs initiatives ont contribué à l'élaboration d'outils qui permettent d'enquêter sur les voies de signalisation et les processus biologiques touchés par un certain composé chimique. Le consortium Gene Ontology (GO) a pris l'initiative de développer une ontologie des gènes qui décrit les processus biologiques, les fonctions moléculaires et les activités biochimiques auxquels le gène contribue. Par conséquent, le concept GO permet, de manière simple et dynamique, l'annotation du gène homologue et les séquences de protéines dans plusieurs organismes en utilisant un vocabulaire commun (Ashburner et al., 2000). En plus de l'ontologie, d'autres approches sont disponibles qui visent à décrire l'interconnexion entre les gènes et les protéines dans un contexte de voie de signalisation biochimique, des exemples sont KEGG (Kanehisa & Goto, 2000) et Reactome (Joshi-Tope et al., 2005).

Un défi important pour appliquer la toxicogénomique était de différencier les médicaments selon leur signature d'expression génique. Initialement, l'objectif ambitieux de plusieurs études était de faire la distinction entre deux classes de médicaments ayant des effets toxiques distincts (médicaments anti-inflammatoires et des agents cytotoxiques endommageant l'ADN), ceci basé uniquement sur la corrélation d'environ 250 profils d'expression de gènes dans les cellules HepG2 humain (Burczynski et al., 2000; Hughes et al., 2000). Hamadeh et al., ont exploité une série d'approches bioinformatiques pour identifier des profils d'expression de gènes à partir du foie de rats mâles Sprague-Dawley, qui distinguaient l'exposition au phénobarbital, un inducteur enzymatique, de ceux exposés aux prolifférateurs de peroxysomes, Clofibrate, gemfibrozil et Wyeth 14.643 (Hamadeh et al., 2002). Par conséquent, l'utilisation des outils d'analyse bioinformatique pour la séparation des classes de composés, basés sur des données d'expression génique, était suffisamment convaincante.

D'autres méthodes plus sophistiquées ont été développées telles que le biclustering (Bergmann, Ihmels, & Barkai, 2003; Csárdi, Kutalik, & Bergmann, 2010; Prelić et al., 2006), cette approche est bien adaptée pour révéler l'organisation modulaire des réponses transcriptionnelles suite à plusieurs perturbations chimiques ce qui est presque toujours le cas dans les grandes études toxicogénomique. Techniquement, chaque bicluster consiste à la fois d'un sous-ensemble de gènes et un sous-ensemble de médicaments (Figure 1.7). Une étude récente a utilisé l'approche biclustering pour générer un grand nombre de modules transcriptionnels induits dans le foie de rat traité par un très grand nombre de médicaments. Cette étude a caractérisé intensivement ces modules de gènes-médicaments en termes de rôles fonctionnels et activités biologiques afin de mieux comprendre les systèmes cellulaires perturbés par la ou les médicaments (Iskar et al., 2013).

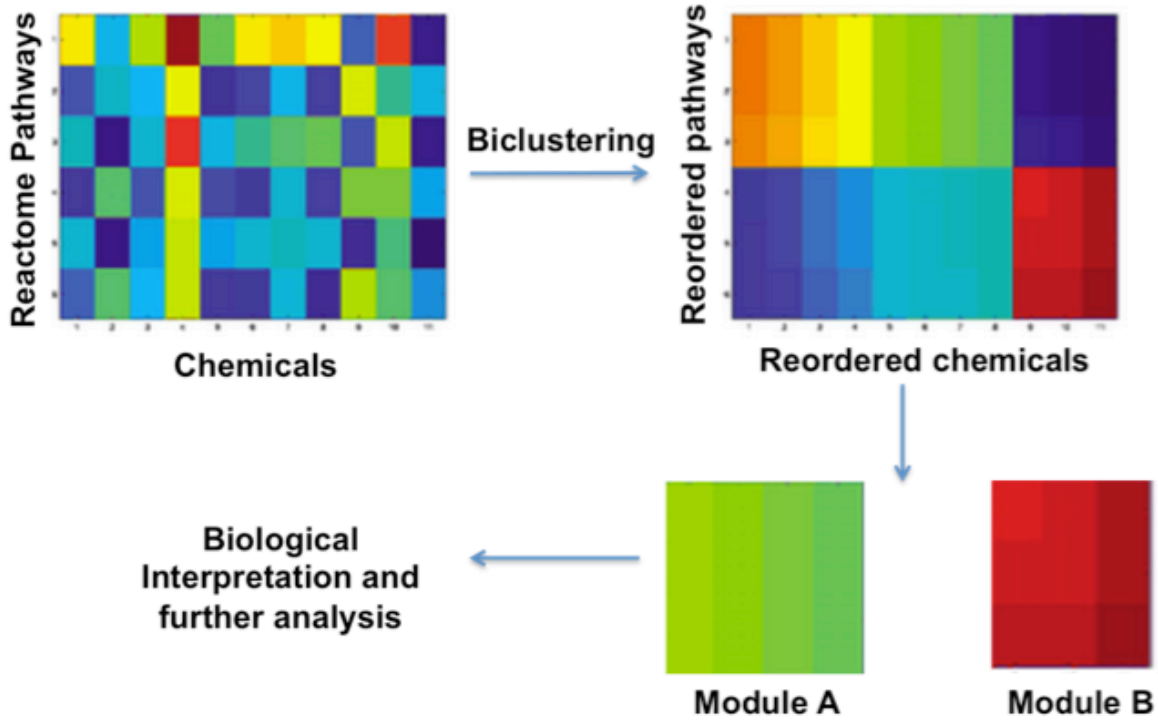

Figure 1.7 cette figure illustre le concept du biclustering. Dans ce type de clustering, l'algorithme génère un motif de gènes/voies de signalisations perturbé par un ensemble de médicaments. Ce motif renseigne sur l'impact biologique de certaines classes de médicaments sur les réseaux moléculaires. Ceci est important pour la détection de biomarqueurs de toxicité.

Une autre analyse récente de TG-GATES a décrit comment l'analyse intégrative de l'expression différentielle des gènes a révélé un ensemble de signatures consensuelles de cytotoxicité chez l'homme et le rat. Ces gènes perturbés forment un réseau fonctionnel conservé qui est responsable de la réponse cellulaire au stress toxique. Enfin, cet ensemble de gènes était prédictif de la toxicité dans un modèle hépatique *in vivo* (Zhang, Berntenis, Roth, & Ebeling, 2014). Cependant, aucune de ces études n'a envisagé une analyse *ab initio* pour identifier les voies des signalisations perturbées par une série de médicaments chez le rat *in vitro* et *in vivo* et dans des hepatocytes humains en culture. Ce type d'études pourra établir une carte des principales voies de toxicité en réponse aux médicaments et perturbateurs environnementaux potentiellement carcinogènes. Ceci sera discuté en détail dans le chapitre 4.

1.6 Rationnelle, hypothèses et objectifs de la thèse

Les progrès de la génomique et des technologies à haut débit tels que les micropuces d'ADN et le séquençage de seconde génération ont conduit à la création de jeux de données diverses, qui se sont avérés très prometteurs pour la médecine de précision et le développement de médicaments. L'analyse de ces ensembles de données pour mieux comprendre le mécanisme d'action des médicaments, identifier des biomarqueurs responsables de la réponse thérapeutique ainsi que la toxicité de nouveaux médicaments est de plus en plus courante. De récents projets scientifiques ont généré et partagé de plus en plus de données moléculaires liées aux expériences sur des modèles biologiques et leur réponse aux médicaments. Nous avons pu utiliser ces grands jeux de données pour mieux comprendre comment un médicament perturbe le système cellulaire en modifiant l'expression de plusieurs gènes cibles. Nous avons aussi pu démontrer qu'un médicament agit efficacement si un marqueur moléculaire est présent dans le contexte cellulaire étudié. En effet, un défi majeur des études menées *in vitro* (lignées cellulaires) est d'extrapoler ces résultats *in vivo* (chez les patients). Ma thèse se focalise sur l'approche centrée sur le médicament (« drug-centric ») et vise à montrer l'utilité et les limitations des modèles cellulaires dans le pipeline du développement du médicament. La disponibilité des grands jeux de données à partir de lignées cellulaires traités par plusieurs classes de médicaments permet d'évaluer la fiabilité de ces modèles à prédire en partie la Réponse thérapeutique ou toxique *in vivo*.

Hypothèses

1) Est-ce que la médiocrité des prédicteurs génomiques de réponse aux antitumoraux est due à une inconsistance dans les mesures de cytotoxicité dans les grandes études pharmacogénomiques publiées?

2) Est-ce que l'intégration de plusieurs types de données à haut débit pour les médicaments (non seulement les antitumoraux), tel que la cytotoxicité, la perturbation du transcriptome ainsi que la structure chimique permet de mieux classifier les médicaments de point de vue taxonomique (mécanisme d'action, indications et cibles)?

3) Est-ce que les perturbations cellulaires toxiques/carcinogènes induits par les médicaments *in vivo*, surtout dans le foie, peuvent être identifiées dans les systèmes *in vitro*?

But principal

L'objectif principal de ma thèse est le développement de méthodes bio-informatiques innovantes pour analyser les nombreuses bases de données pharmaco — et toxicogénomiques publiques afin de comprendre la réponse thérapeutique aux médicaments anticancéreux (chapitre 2), la similarité des mécanismes d'action-cibles thérapeutiques des médicaments (chapitre 3) et les réponses toxicogénomiques relatives aux médicaments et autres produits toxiques (chapitre 4). Les données pharmaco — et toxicogénomiques seront tout d'abord collectées et organisées. Nous nous focaliserons sur les données d'expressions géniques (transcriptomique), car ces données sont disponibles pour les études de cytotoxicité, perturbations et d'hépatotoxicité. Nous analyserons ces grands jeux de données pour valider ou invalider nos trois hypothèses de travail.

Comme illustré dans la Figure 1.1, les dénominateurs communs à tous les chapitres de cette thèse sont :

- les profils d'expressions de gènes produits par de grandes études de pharmacogénomique et toxicogénomique.
- Les médicaments (molécules chimiques déjà approuvées ou encore expérimentales)

- Les lignées cellulaires (lignées cancéreuses pour chapitre 2, 3 et lignée hépatique, chapitre 4)

Le chapitre 2 comprend une étude comparative entre CCLE et GDSC pour mieux caractériser la concordance/discordance au niveau des mesures génomiques et profils cellulaires de sensibilité aux médicaments. Ceci est d'une grande importance pour identifier des biomarqueurs de réponse aux anticancéreux, un projet ambitieux pour la médecine personnalisée.

Le chapitre 3 comprend une étude intégrative de plusieurs types de données (pharmaco) génomiques (CTRPv2, NCI60, LINCS) pour mieux caractériser le mécanisme d'action de médicaments sans nécessairement avoir recours aux données *a priori* disponible pour le médicament, c'est-à-dire la cible biologique ou indication clinique.

Le chapitre 4 comprend une étude intégrative de la base de données TG-GATES pour caractériser les modules transcriptionnels conservés, et induits par une variété de classes de médicaments, *in vitro* et *in vivo* chez le rat et l'humain (seulement *in vitro*). Notre étude a permis de montrer que le modèle *in vitro* est capable d'identifier des modules pertinents liés à l'hépatotoxicité et la carcinogenèse chimique.

1.7 Organisation de la thèse

Chapitre 1 : Le premier chapitre est une brève revue de la bibliographie permettant à un lecteur non-spécialiste d'avoir une vue d'ensemble sur les concepts et approches utilisées dans cette thèse.
Chapitre 2 à 4 : Les principaux résultats de la thèse présentés sous forme d'articles pour revues scientifiques.
Chapitre 5 : Ce chapitre présentera une discussion sur l'apport et l'impact de ma recherche à l'étude des données de pharmaco/toxicogénomique.

1.8 Reproductibilité de la recherche

Le Dr Haibe-Kains porte une grande attention sur la réplicabilité de la recherche effectuée au sein de son laboratoire. Je ne peux qu'appuyer cette démarche qui est hautement importante en recherche computationnelle où il devrait être aisé d'obtenir les mêmes résultats en refaisant tourner le pipeline d'analyse (Sandve, Nekrutenko, Taylor, & Hovig, 2013). Au cours de ma thèse, j'ai donc adhéré aux principes de la recherche reproductible en décrivant de manière adéquate mes méthodes (notamment dans les informations supplémentaires de mes articles), en partageant publiquement les données et codes nécessaires pour reproduire toutes les figures, tables et autres résultats qui font partie de mes publications. Si nécessaire, un site compagnon fut créé pour fournir des fichiers et descriptions additionnels concernant mes études (pmgenomics.ca/bhklab/reproducibility).

CHAPITRE 2

INCONSISTENSY IN LARGE PHARMACOGENOMIC STUDIES

L'un des principaux objectifs des études de lignées cellulaires cancéreuses est de tester l'efficacité des agents thérapeutiques (médicaments) et ainsi déterminer les facteurs génomiques prédictifs de la réponse aux anticancéreux dans le but de maximiser l'efficacité et réduire les effets secondaires néfastes.

Deux grandes études de pharmacogénomiques ont été publiées en 2012. Les ensembles de données obtenus présentaient une occasion unique pour caractériser les caractéristiques génomiques/transcriptomiques associées à la réponse aux médicaments. Notre anayse a révélé que ces deux études ont testé 471 lignées cellulaires et 15 médicaments en commun. Nos résultats comparatifs ont montré que l'expression des gènes est bien corrélée entre les études pour les mêmes lignées cellulaires, alors que les réponses pharmacologiques sont très discordantes pour le même médicament. Ceci est surprenant puisque les deux études ont utilisé la mesure IC_{50} (concentration du médicament qui inhibe 50 % de la croissance cellulaire) et AUC (aire sous la courbe présentant la viabilité cellulaire en fonction de la concentration du médicament). Bien que la source de discordance demeure incertaine, notre étude a montré que la standardisation des essais pharmacologiques est importante pour la collecte de biomarqueur génomique reproductible et ainsi valider ultérieurement ces données dans des études précliniques et cliniques.

Contributions par auteur :

NEH a collecté, préparé, organisé les données de pharmacogénomique, a contribué au code, au concept de réseaux de gènes (« pathways ») et multi biomarqueurs, à l'interprétation des résultats ainsi qu'à la rédaction du manuscrit. AJC et AHB ont aidé à l'analyse des données de mutations. BHK a conçu le design de l'analyse, écrit le code en collaboration avec NEH et

supervisé l'étude. NJB, AHB, HJWLA et JQ ont contribué à l'interprétation des résultats et à la rédaction du manuscrit.

Cette étude a été publiée dans le journal *Nature* :

*Haibe-Kains B, **El-Hachem N**, Birkbak NJ, Jin AC, Beck AH, Aerts HJ, Quackenbush J.* Inconsistency in large pharmacogenomic studies. *Nature*. 2013

Inconsistency in large pharmacogenomic studies

Benjamin Haibe-Kains[1, 2, $], Nehme El-Hachem[1], Nicolai Juul Birkbak[3], Andrew C. Jin[4], Andrew H. Beck[4,*], Hugo J.W.L. Aerts[5, 6,7,*], John Quackenbush[5.8,*]

[1]Institut de Recherches Cliniques de Montréal, University of Montréal, Montréal, Québec, Canada; [2]Ontario Cancer Institute, Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada; [3]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark; [4]Department of Pathology, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, MA, USA; [5]Department of Biostatistics and Computational Biology and Center for Cancer Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA; [6]Department of Radiation Oncology & Radiology, Dana-Farber Cancer Institute, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA; [7]Department of Radiation Oncology, Maastricht University, Maastricht, The Netherlands; [8]Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA

*Co-last authors

[$] Corresponding author : Benjamin Haibe-Kains bhaibeka@uhnresearch.ca

2.1 Abstract

Cancer cell line studies have long been used to test efficacy of therapeutic agents and to explore genomic factors predictive of response [1-2]. Two large-scale pharmacogenomic studies were published recently [3-4]; each assayed a panel of several hundred cancer cell lines for gene expression, copy number, genome sequence, and pharmacological response to multiple anti-cancer drugs. The resulting datasets present a unique opportunity to characterize mechanisms associated with drug response, with 471 cell lines and 15 drugs assayed in both.

However, while gene expression is well correlated between studies, the measured pharmacologic drugs response is highly discordant. This poor correspondence is surprising as both studies assessed drug response using common estimators: the IC_{50} (concentration at which the drug inhibited 50% of the maximal cellular growth), and the AUC (area under the activity curve measuring dose response)[5]. For drugs screened in both studies, only one had a Spearman correlation coefficient in measured response greater than 0.6. Importantly these results are also reflected in inconsistent associations between genomic features and drug response. Although the source of inconsistencies in drug response measures between these two well-controlled studies remains uncertain, it makes drawing firm conclusions about response very difficult and has potential implications for using these outcome measures to assess gene-drug relationships or select potential anti-cancer drugs based on their reported results. Our findings suggest standardization of response measurement protocols in pharmacogenomic studies is essential before such studies can live up to their promise.

2.2 Methods

To ensure reproducibility of our analysis, we developed an automated pipeline in R that can generate all the results, figures and tables of the paper (see extended figures and Supplementary data in the online version for full details).

2.2.1 Data retrieval and curation

We retrieved and curated data from three large pharmacogenomic studies, namely the Cancer Genome Project (CGP), the Cancer Cell Line Encyclopedia (CCLE) and the GlaxoSmithKline cell line collection.

For CGP, gene expression data (raw Affymetrix CEL files) were downloaded from ArrayExpress (http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-783/). Drug sensitivity measurements, mutation data and cell lines annotations were downloaded from the CGP website (http://www.cancerrxgene.org/downloads/). Drug information was collected from Supplementary Information of Garnett *et al.[4]*. Minimum and maximum screening

concentrations (μM) for each drug/cell line were extracted from gdsc_compounds_conc_w2.csv available on the CGP website. The natural logarithm of IC_{50} measurements were retrieved from column "*_IC_50" of gdsc_manova_input_w2.csv available on the CGP website. The AUC measurements were retrieved from gdsc_manova_input_w2.csv in column "*_AUC". Coding variants in 68 genes were also extracted from gdsc_manova_input_w2.csv.

For CCLE, gene expression, mutation data cell line annotations and drug information were downloaded from the CCLE website (http://www.broadinstitute.org/ccle) Drug sensitivity data were downloaded from the addendum published by Barretina *et al*.[6]. Screening concentrations (μM) for each drug/cell line were extracted from column "Doses (μM)". IC_{50} measurements were retrieved from column "IC_{50} μM (norm) ». AUC measurements were retrieved in column "ActArea (norm) ». Coding variants in 1667 genes (column 'Protein Change') measured using the Oncomap3 and hybrid capture platforms were extracted from CCLE_Oncomap3_2012-04-09.maf an CCLE_hybrid_capture1650_hg19_NoCommonSNPs_NoNeutralVariants_CDS_2012.05.07.maf, respectively.

For GSK, gene expression data and cell line annotations were downloaded from the National Cancer Informatics Program website (http://cbiit.nci.nih.gov/ncip). IC_{50} measurements and drug information were downloaded from Supplementary Table 2 (stab_2.xls) of Greshock *et al*. [7].

2.2.2 Cell line annotation

Cell line names were harmonized in CGP, CCLE and GSK to match identical cell lines; this was done through manual search over alternative names of cell lines, as reported in the corresponding cell line annotation files and online databases such as hyperCLDB (http://bioinformatics.istge.it/hypercldb/) and BioInformationWeb (http://bioinfoweb.com). We identified 471 cancer cell lines being investigated both in CGP and CCLE, 231 cell lines shared between CGP and GSK, 249 cell lines shared between CCLE and GSK, and 194 cell lines shared by all three studies (Figure II.Ic). To annotate the tissue of origin of each cell lines

we chose the nomenclature used in CGP; CCLE and GSK tissue type information was therefore updated to follow this nomenclature, which resulted in 24 tissue types.

2.2.3 Drug sensitivity data

Drug sensitivity measures, which are IC_{50} and AUC values, were set to common scale ($-\log_{10}(\mathrm{M})$ for IC_{50} and [0,1] for AUC) across studies so that high values are representative of cell line sensitivity to drugs. For CGP, extracted IC_{50} measures (*x*) were transformed using $-\log_{10}(\exp(x)/10^6)$, and AUC measures were left untransformed. For CCLE, extracted IC_{50} measures (*x*) were transformed into logarithmic scale, $-\log_{10}(x/10^6)$, and AUC measures were divided by the number of drug concentrations tested. For GSK, extracted IC_{50} measures (*x*) were transformed using $-\log_{10}(x/10^3)$.

We also discretized the drug sensitivity measures into three categories (resistant, intermediate and sensitive) using the waterfall method described in the CCLE study[3]. The full procedure, as provided by Dr. Kavitha Venkatesan (personal communication) is described below:

1. Extract the drug sensitivity measurements, either IC_{50} or AUC.
2. Sort increasing log IC_{50} values (or AUC) of the cell lines to generate a waterfall distribution.
3. If the waterfall distribution is non-linear (Pearson correlation coefficient to the linear fit $\leq$ 0.95), estimate the major inflection point of the log IC_{50} curve as the point on the curve with the maximal distance to a line drawn between the start and end points of the distribution.
4. If the waterfall distribution appears linear (Pearson correlation coefficient to the linear fit > 0.95), then use the median IC_{50} instead.
5. Cell lines within a 4-fold IC_{50} (or within a 1.2-fold AUC) difference centered around this inflection point are classified as being intermediate, cell lines with lower IC_{50} (or AUC) values than this range are defined as sensitive, and those with IC_{50} (or AUC) values higher than this range are called resistant.
6. Require at least *x*=5 sensitive and *x*=5 resistant cell lines after applying these criteria.

Using this approach we generated drug sensitivity calls for all drugs in CGP and CCLE.

2.2.4 Gene expression data

Raw gene expression profiles (Affymetrix CEL format) for 789 CGP, 1036 CCLE and 950 cell lines were downloaded, respectively, from ArrayExpress[8] (E-MTAB-783), CCLE (www.broadinstitute.org/ccle/) and NCIP (http://cbiit.nci.nih.gov/ncip) websites. Gene expression data were normalized with frozen RMA[9] using the Bioconductor Chip Description File (CDF) definitions (hthgu133a for CGP, and hgu133plus2 for CCLE and GSK, respectively). We then used the R package *jetset[10]*, which maps Affymetrix probe sets to unique Entrez gene ids by selecting the best probe set for each gene; subsequent analyses were restricted to the 12,187 probe sets common to the CGP, CCLE and GSK arrays. For replicates in CGP and GSK, the CEL files were ordered by hybridization date and the first experiment was selected.

2.2.5 Mutation data

Missense mutations in 64 protein-coding genes sequenced in 431 cell lines both in CGP and CCLE were downloaded from their respective website. Similarly to CGP and CCLE studies [3-4], mutation data were discretized to represent the presence or absence of missense mutation in a given gene in a given cell line.

2.2.6 Gene-drug associations

We assessed the association between gene expression and drug response, referred to as gene-drug association, using a linear regression model controlled for tissue source:

$$Y = \beta_0 + \beta_i G_i + \beta_t T$$

where Y denote the drug sensitivity variable, G_i and T denote the expression of gene i and the tissue type respectively, and βs are the regression coefficients. The strength of gene-drug association is quantified by β_i, above and beyond the relationship between drug sensitivity and

tissue source. The variables Y and G are scaled (standard deviation equals to 1) to estimate standardized coefficients from the linear model. Significance of the gene-drug association is estimated by the statistical significance of β_i (two-sided t test).

2.2.7 Pathway-drug associations

For each drug, genes were ranked according to the statistical significance of their gene-drug association (Student t statistic). We then used this drug-specific gene ranking to perform pre-ranked gene set enrichment analyses (GSEA[11] version 2.0.13) in order to assess enrichment of gene ontology terms[12] curated in MSigDB[11] (c5.all.v4.0.entrez.gmt). Only pathways whose corresponding gene sets contained between 15 genes and 250 genes, were considered for further analysis (913 gene sets). We used the resulting normalized enrichment (NES) scores to quantify the strength of pathway-drug associations.

2.2.8 Measures of consistency

We computed Spearman rank-ordered correlation coefficients (r_s) [13] to assess the consistency between CGP and CCLE drug phenotypes (IC_{50} and AUC measures), gene/mutation-drug associations (coefficient β) and pathway-drug associations (normalized enrichment scores). We used Cohen's Kappa (κ) coefficient [14] to assess consistency between CGP and CCLE drug sensitivity calls (resistant, intermediate, sensitive) and mutation data. We used the following qualitative descriptions of correlation coefficient (r_s) values associated with intervals: $r_s < 0.5$, poor consistency; $0.5 \leq r_s < 0.6$, fair consistency; $0.6 \leq r_s < 0.7$, moderate consistency; $0.7 \leq r_s < 0.8$, substantial consistency; and $r_s \leq 0.8$, almost perfect consistency. Same qualitative descriptions were used for Cohen's Kappa (κ) coefficient.

2.3 Background and Results

Patients with cancer often exhibit heterogeneous responses to anticancer treatments and evidence suggests response is determined in part by patient-specific alterations in the somatic cancer genome and changes in gene expression[15]. A number of studies have searched for gene expression signatures predictive of response, however most only tested a limited number of genes, a small panel of drugs, or assayed drug response in a small number of cell lines [1][16-17].

Results from two large-scale pharmacogenomic studies, the Cancer Genome Project (CGP)[4] and the Cancer Cell line Encyclopedia (CCLE)[3], were recently reported in this journal. The CGP tested 138 anti-cancer drugs against 727 cell lines while the CCLE tested response of 24 drugs against 1036 cell lines (Figure II.I); of thèse, 15 drugs (Figure II.Ia, b) and 471 cell lines were tested in both (Figure II.Id, e). Both groups tested mutations in 64 genes (Figure II.Ig) and expression of 12,153 genes (Figure II.Ih) genes. The overlap allows assessment of consistency between these independent datasets and the potential to infer genomic models predictive of drug response.

We downloaded, curated, and annotated the genomic and pharmacological data from the CGP and CCLE studies (Methods). We first compared expression profiles between the 61 biological replicates in CGP and observed very high correlation (median Spearman correlation of 0.97; Figure 1a) indicating excellent reproducibility within the same study.

We then compared gene expression profiles of the 471 cell lines shared between studies. Despite the use of different array platforms (Affymetrix GeneChip HG-U133A in CGP and HG-U133PLUS2 in CCLE), the expression profiles of identical cell lines were significantly better correlated than between different cell lines (median correlation of 0.85 vs. 0.34 for identical and different cell lines, respectively; two-sided Wilcoxon Rank Sum test p-value $< 1\times10^{-16}$). For 467 cell lines, the most highly correlated gene expression profile was with the same cell line; only four (MOG-G-CCM, SNB19, SW1990, and SW403) were more highly correlated with another cell line (Figure 2.1 b). This small discordance between the CGP and CCLE is likely due to experimental artifacts, measurement error, or divergence of the four cell lines. We tested consistency based on the tissue from which the cell line was derived (Supplementary Figure 2.1). We found the highest correlation, with cell lines from the urinary tract (median correlation of 0.87) and the lowest for those the upper aerodigestive tract (median correlation of 0.79),

We compared the reported presence of mutations for 64 genes in the shared 471 cell lines and found better agreement between identical cell lines than between different cell lines (two-sided Wilcoxon Rank Sum test p-value < 1×10^{-16}; Figure II.II), although not perfect agreement (median Cohen's Kappa [**κ**] of 0.65), which might be due to the different sequencing platforms and software used to call genomic variants in the two studies. Agreement in mutation profiles was higher in pancreas cell lines although the difference was not significant (Supplementary Figure 2.2).

We then compared drug sensitivity phenotype measurements. In the CGP drug screening was performed at two sites, the Massachusetts General Hospital (MGH) and the Wellcome Trust Sanger Institute (WTSI). As a control, Camptothecin, an inhibitor of DNA enzyme topoisomerase I, was screened at both sites using the same experimental protocol in 252 cell lines. The IC_{50} (concentration in micro molar [μM] at which the drug inhibited 50% of the maximum cellular growth) for Camptothecin had significant but only fair correlation (r_s=0.58. p-value=$1.5\text{x}10^{-23}$, Figure II.III).

We compared drug sensitivity measures between CGP and CCLE in fifteen drugs (Figure II.Ia,b) tested on the 471 shared cell lines (Figure II.Id,e). Both CGP and CCLE measured cell line drug sensitivity using IC_{50} and AUC (area under the activity curve measuring dose response), also referred to as Activity Area [5]; however the two studies used different experimental protocols (summarized in Supplementary Information online). Differences include the pharmacological assay used, the range of drug concentrations tested, and choice of an estimator for summarizing the drug dose-response curve.

In both studies, the IC_{50} could not be estimated in many cases, as drug concentration necessary to inhibit 50% of growth was not reached. In CGP, IC_{50} was estimated using a Bayesian sigmoid model for drug response. In contrast, CCLE reported the maximum concentration for inactive compounds (referred to as placeholder values) rather than the extrapolated IC_{50} . AUC measures do not require extrapolation and can always be estimated from the dose response curve.

For each of the 15 drugs assayed by both CGP and CCLE we ranked the response of the 471 shared cell lines (Figure 2a) and computed the Spearman correlation coefficient (see Methods) for the reported IC_{50} (Figure 2.2b). We found a single drug, 17AAG (an HSP90

inhibitor), with moderate correlation (r_s=0.61; Table II.Ia) and another, PD0325901 (a MEK inhibitor), with fair correlation (r_s=0.53; Table II.Ia) between studies.

To test whether extrapolation decreased the correlations between studies we filtered out all IC_{50} values exceeding the maximum tested drug concentrations. We observed only small increases in correlation for PLX4720, PD0325901 and Paclitaxel and decreases for 17AAG and AZD6244, although the number of measurements was small (Figure II.IV). We also compared reported AUC measures (Figure 2.2b, Table II.Ib, Figure II.V) and found that only two drugs yielded fair correlations (17AAG with r_s=0.58 and PD0325901 with r_s=0.55).

We compared correlations computed from AUC and IC_{50} (Figure 2.2b) and found AUC is more concordant between CGP and CCLE (median correlation of 0.35 and 0.28 for IC_{50} and AUC, respectively) but that the difference was not significant (two-sided Wilcoxon signed rank test p-value=0.3). The vast majority of drugs yielded poor concordance (r_s<0.5) for both IC_{50} and AUC, which suggest that the lack of consistency of the drug response cannot be solely explained by the choice of the estimator of drug sensitivity.

We tested whether drug response correlation depended on tissue source. We found both IC_{50} and AUC measures tend to be more consistent in cell lines originating from urinary tract; this difference is significant for AUC (two-sided Kruskal-Wallis test p-value=0.024). However, due to the small number of urinary tract cell lines (10), only three drugs (PD0325901, Nutlin-3 and 17AAG) had statistically significant moderate correlation.

In addition to IC_{50} and AUC, we also compared sensitivity using the *waterfall* method described in the CCLE study. Drug sensitivity calls (resistant, intermediate and sensitive) were estimated from IC_{50} and AUC values and compared using Cohen's $\boldsymbol{\kappa}$ (see Methods). Again, the drug sensitivity calls for both IC_{50} and AUC estimates had a poor agreement between studies ($\boldsymbol{\kappa} < 0.5$).

Despite the discordance in drug sensitivity measures between CGP and CCLE, we tested whether the association between drug response and genomic features might be consistent across datasets. This is important because the identification of genomic predictors of drug response was the primary goal of both the CGP and CCLE studies.

We estimated gene-drug associations by fitting, for each gene, a linear regression model including gene expression as predictor of drug sensitivity, controlled for tissue source (see Methods). Linear models were fitted using both IC_{50} and AUC measures. Here too, we

observed poor correspondence between studies, the best correlation with IC_{50} data was observed for 17AAG (r_s=0.38; Figure 2.3a, and Table II.Ia); for the vast majority of drugs correlations were slightly better when AUC measures were used to estimate gene-drug associations but the best correlation was still poor (r_s=0.46 for PD0325901; Figure 2.3a, Table II.Ib and Figure II.VI). Although correlations significantly depended on tissue sources (two-sided Kruskal-Wallis test p-value < 0.006), only drugs screened in hematopoietic/lymphoma tissue and urinary tract yielded slightly higher correlation than all tissues combined for both IC_{50} and AUC.

We tested whether these poor correlations could be due to genes unrelated to drug sensitivity by focusing on genes statistically associated with drug sensitivity (false discovery rate, FDR <20%) in at least one dataset. Overall, while the correlations were better than those computed using all genes, they were still low. For $IC_{50,}$ only AZD6244 and 17AAG yielded a moderate correlation (r_s=0.65 and r_s=0.63, respectively; Table II.Ia). Using AUC and this subset of genes, we found that PD0332991 had fair correlation, and five drugs had moderate correlation between studies (PD0325901, AZD6244, Nilotinib, 17AAG, and Nutlin-3; Table II.Ib). However the correlations for the remaining drugs remained poor and did not significantly depend on tissue source (two-sided Kruskal-Wallis test p-value > 0.064).

We recognize that activation of drug-response through specific gene functional classes may be more predictive than individual genes. We therefore used the previously computed gene-drug associations to rank genes by the significance of their association with drug sensitivity and searched for over-represented Gene Ontology (GO) terms using pre-ranked gene set enrichment analysis (GSEA). We compared the normalized enrichment scores computed for CGP and CCLE for the 15 drugs screened in both studies (see Methods).

For IC_{50}, there was poor correlation of GSEA enrichment scores for drugs, except for AZD6244 and PD0325901, which yielded fair correlation (r_s=0.63 for AZD6244 and r_s=0.68 for PD0325901; Figure 2.3c, Table II.Ia). When using AUC, two drugs yielded fair correlations (Nilotinib, 17AAG), AZD6244 yielded moderate correlation and PD0325901 yielded substantial correlation (r_s=0.76; Figure 2.3c, Table II.Ib, Figure II.VII). These correlations significantly depended on tissue source (two-sided Kruskal-Wallis test p-value < $7x10^{-4}$) where median drug correlations computed from IC_{50} were higher in breast, urinary tract, hematopoietic/lymphoma and lung cell lines compared to all tissues combined.

We repeated the analyses, this time focused on the GO classes that are statistically significantly enriched (FDR < 20% for normalized enrichment score) among genes associated with drug response in at least one of the two studies. Using IC_{50}, most correlations increased slightly, except for 17AAG and PD0332991, with PLX4720 and PD0325901 yielding moderate correlation (Figure 2.3c and 2.3d, Table II.Ia). For AUC, we observed fair correlation for Paclitaxel and Sorafenib, moderate correlation only for Lapatnib, and substantial correlation for PD0325901 and AZD6244 (Figure 2.3d, Table II.Ib).

These pathway-based correlations are the best observed in our analysis as almost half of the drugs exhibited a correlation greater than 0.5, although they are still quite poor. When stratifying by tissue source, only drugs screened in lung cancer cell lines yielded slightly higher median correlation compared to all tissues combined.

We then performed similar analyses using mutation data of the 64 genes sequenced both CGP and CCLE (Figure II.Ig). We observed that few mutations were significantly associated with drug response, which partly explains the poor correlation between mutation-drug associations ($r_s < 0.5$; Figure II.VIII).

To test whether genomic data or drug response measures are the likely source of the poor correlations, we used identical (therefore perfectly correlated) gene expression data for the 471 cell lines while keeping the original drug sensitivity measures in each study, but did not find improved correlations for (significant) gene-drug associations (see 'GeneCGP fixed' and 'GeneCCLE fixed' in Figure 2.4). However when using identical drug phenotypes with the original gene expression data, correlations significantly increased in all cases (two-sided Kruskal-Wallis test p-value < 0.01, see 'DrugCGP fixed' and 'DrugCCLE fixed' in Figure 2.4) and yielded almost perfect correlation for significant gene-drug associations with AUC (median correlation > 0.83). Results were similar for pathway-drug associations. These results clearly demonstrate that the discordance between studies stems from the drug sensitivity measurements.

We also investigated the impact of the choice of pharmacological assay across study and compared CGP and CCLE drug sensitivity data with those published by Greshock and colleagues in a panel of 319 cell lines; the GlaxoSmithKline (GSK) dataset. The GSK authors used the same pharmacological assay used by the CCLE (Cell Titer Glo Luminescent Cell Viability Assay kit from Promega), but other parameters in the experimental protocols differ

from those in either CGP or CCLE and they used yet another model to estimate IC_{50} values (model 205 in XLfit in Microsoft Excel).

Among the fifteen drugs shared between CGP and CCLE, only two, Lapatinib and Paclitaxel, were tested by GSK on a common set of 194 cell lines. As might be expected based on the assay used, GSK IC_{50} measurements were more consistent with those of CCLE IC_{50} (r_s=0.42 and 0.36 for Lapatinib and Paclitaxel, respectively) than CGP (r_s=0.24 and 0.10 for Lapatinib and Paclitaxel, respectively), but here too the overall consistency was rather poor (and similar to the observed consistency between CCLE and CGP).

We then performed the same analysis but focusing on drugs and cell lines shared only by two studies. For Lapatinib and Paclitaxel, screened by CCLE and GSK in 249 common cell lines, we observed fair to poor correlations (Figure II.IXa). Five drugs and 231 cell lines were screened both in CGP and GSK (Figure II.Ic); for these we observed poor correlation (r_s ranging from 0.12 to 0.30; Figure II.IXb).

These results add further evidence that the inconsistency between studies stems from the use of different pharmacological assays, but there is no clear evidence to conclude which of the three approaches is more accurate. Indeed, even if we observed perfect correlation between GSK and either the CGP or CCLE drug response assays, all that would indicate is a consistency in measurement, but not necessarily which provided the most meaningful assay of drug response or which could best be translated to *in vivo* response.

Our analysis of these three large-scale pharmacogenomic studies points to a fundamental problem in assessment of pharmacologic drug response. While gene expression analysis has long been seen as a source of "noisy" data, extensive work has led to standardized approaches to data collection and analysis and the development of robust platforms for measuring expression levels. This standardization has led to substantially higher quality, more reproducible expression data sets, and this is evident in the CCLE and CGP data where we found excellent correlation between expression profiles in cell lines profiled in both studies.

The poor correlation between drug response phenotypes is troubling and may represent a lack of standardization in experimental assays and data analysis methods. However, there may be other factors driving the discrepancy. As reported by the CGP, there was only a fair correlation ($r_s < 0.6$) between Camptothecin IC_{50} measurements generated at two sites using matched cell line collections and identical experimental protocols. While this might lead to

speculation that the cell lines could be the source of the observed phenotypic differences, this is highly unlikely as the gene expression profiles are well correlated between studies.

While our analysis has been limited to common cell lines and drugs between studies, it is not unreasonable to assume that the measured pharmacogenomic response for other drugs and cell lines assayed are also questionable. Ultimately, the poor correlation in these published studies presents an obstacle to using the associated resources to build or validate predictive models of drug response. Because there is no clear concordance, predictive models of response developed using the data from one study are almost guaranteed to fail when validated on data from the other [18] and there is no way with available data to determine which study is more accurate. This suggests that users of both datasets should be cautious in their interpretation of results derived from their analyses.

Clearly the investment in these projects warrants additional work to resolve the discrepancies in drug response phenotype so that the wealth of data that has been generated can be used to its fullest advantage. Our findings support the need for standardization of drug-response measurements or development of new, robust drug sensitivity assays; without such assays, it will not be possible to reliably identify genomic predictors of drug response or effectively a drug's mechanism of action.

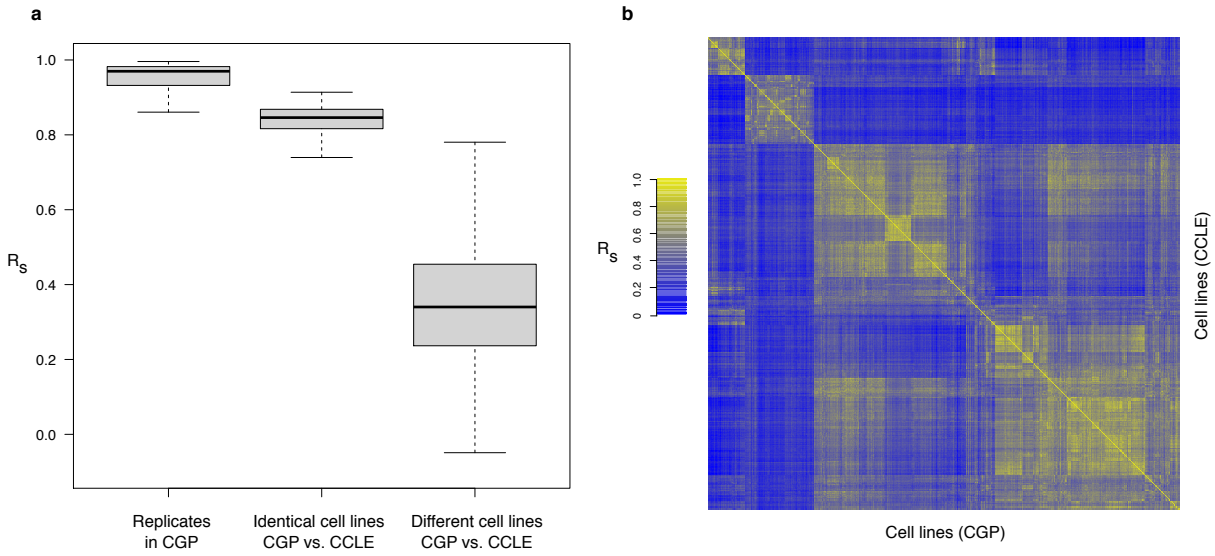

Figure 2.1 : Consistency between gene expression profiles of cell lines in CGP and CCLE studies. (a) Box plot representing the correlation coefficients of the biological replicates in CGP, identical and between different cell lines from CGP and CCLE datasets; (b) heatmap representing the correlations between gene expression profiles of cell lines; the order of cell lines is identical in rows (CCLE) and columns (CGP).

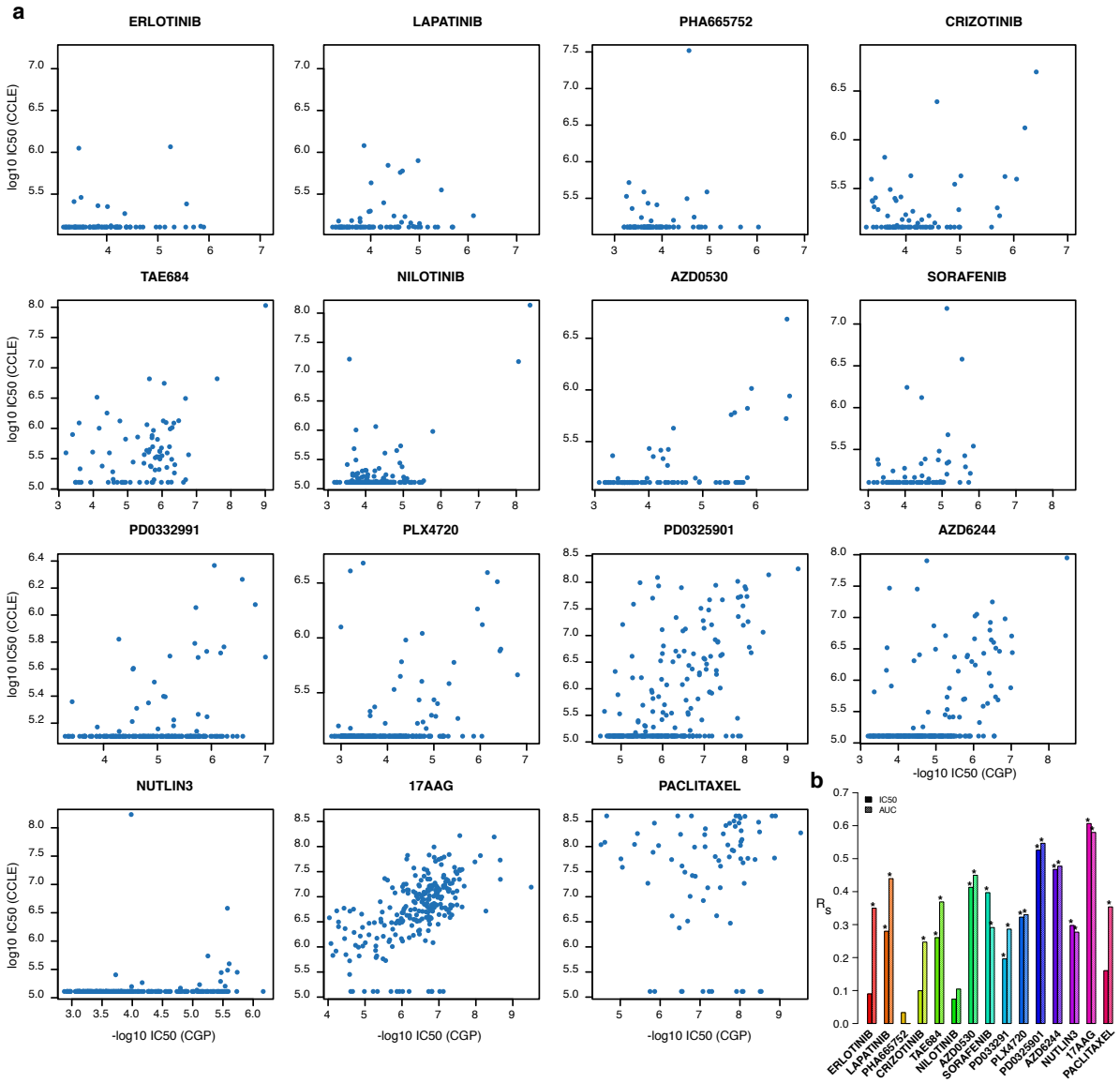

Figure 2.2 : Consistency between drug sensitivity data published in CGP and CCLE studies. (a) Scatter plots reporting the drug sensitivity (IC_{50}) measured in the 471 cell lines and for the 15 drugs investigated both in CGP and CCLE. (b) Bar plot representing the Spearman correlation coefficient for IC_{50} and AUC drug sensitivity measures; significance is reported using the symbol '*' if two-sided p-value < 0.05.

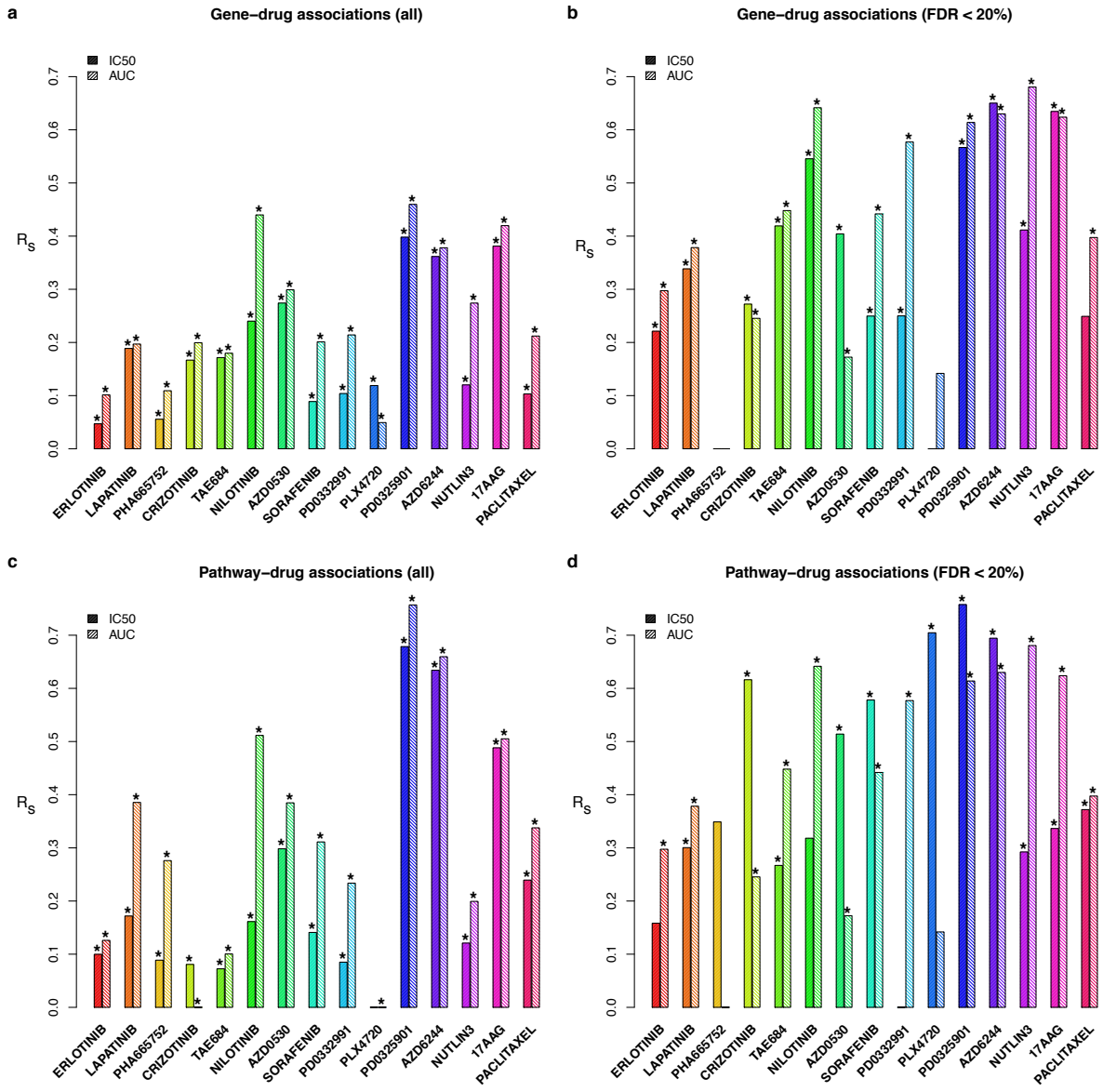

Figure 2.3 : Consistency of associations of genomics features with drug sensitivity. The bars represent the Spearman correlation coefficients computed from: (a) all and (b) significant (FDR<20%) gene-drug associations; (c) all and (d) significant (FDR<20%) pathway-drug associations, as estimated in CGP and CCLE datasets. Significance is reported using the symbol '*' if two-sided p-value < 0.05.

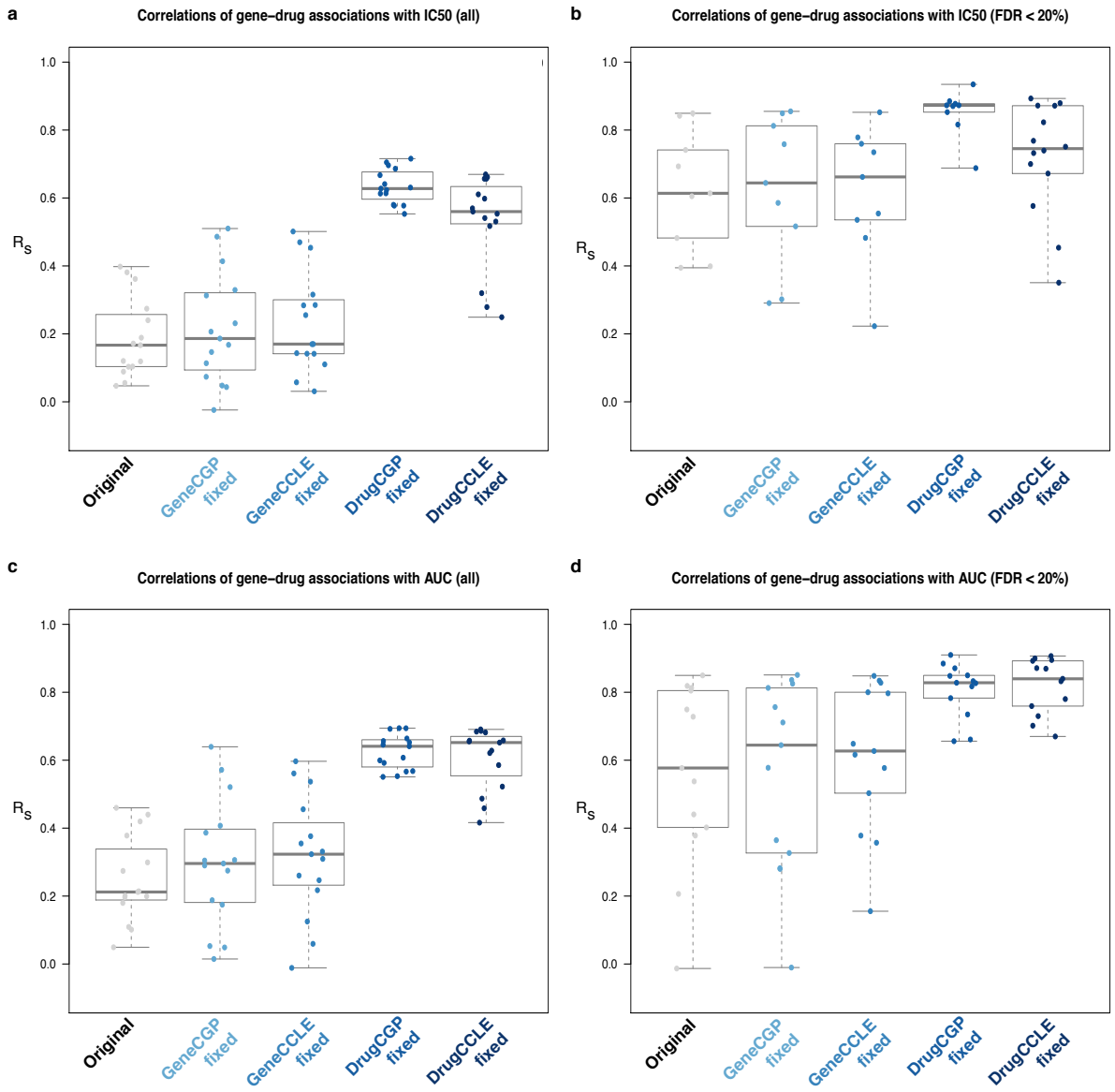

Figure 2.4 : Effects on consistency by intermixing CCLE and CGP data. The box plots report the correlations between: (a) all and (b) significant (FDR < 20%) gene-drug associations with IC_{50}; (c) all and (d) significant (FDR < 20%) gene-drug associations with AUC. Each box represent the datasets used to compute correlations: 'Original' refers to the original datasets; » GeneCGP.fixed » refers to [CGP_g+CGP_d] vs. [CGP_g+$CCLE_d$]; » GeneCCLE.fixed » refers to [$CCLE_g$+CGP_d] vs. [$CCLE_g$+$CCLE_d$]; » DrugCGP.fixed » refers to [CGP_g+CGP_d] vs. [$CCLE_g$+ CGP_d]; » DrugCCLE.fixed » refers to [CGP_g+$CCLE_d$] vs. [$CCLE_g$+$CCLE_d$] where $_g$ and $_d$ stand for gene expression and drug sensitivity data, respectively.

1. Shoemaker, R. H. The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer* **6,** 813–823 (2006).
2. Weinstein, J. N. Drug discovery : Cell lines battle cancer. *Nature* **483,** 544–545 (2012).
3. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–607 (2012).
4. Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483,** 570–575 (2012).
5. Wu, R. & Lin, M. *Statistical and computational pharmacogenomics*. (CRC Press, 2008).
6. Barretina, J. *et al.* Addendum : The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **492,** 290–290 (2012).
7. Greshock, J. *et al.* Molecular target class is predictive of in vitro response profile. *Cancer Res.* **70,** 3677–3686 (2010).
8. Parkinson, H. *et al.* ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.* **35,** D747–50 (2007).
9. McCall, M. N., Bolstad, B. M. & Irizarry, R. A. Frozen robust multiarray analysis (fRMA). *Biostatistics* **11,** 242–253 (2010).
10. Li, Q., Birkbak, N. J., Gyorffy, B., Szallasi, Z. & Eklund, A. C. Jetset : selecting the optimal microarray probe set to represent a gene. *BMC Bioinformatics* **12,** 474 (2011).
11. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102,** 15545–15550 (2005).
12. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25,** 25–29 (2000).
13. Spearman, C. The proof and measurement of association between two things. By C. Spearman, 1904. *Am. J. Psychol.* **100,** 441–471 (1987).
14. Sim, J. & Wright, C. C. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys. Ther.* **85,** 257–268 (2005).
15. Roden, D. M. & George, A. L., Jr. The genetic basis of variability in drug responses. *Nat. Rev. Drug Discov.* **1,** 37–44 (2002).
16. Heiser, L. M. *et al.* Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 2724–2729 (2012).

17. Yamori, T. Panel of human cancer cell lines provides valuable database for drug discovery and bioinformatics. *Cancer Chemother. Pharmacol.* **52 Suppl 1,** S74–9 (2003).
18. Papillon-Cavanagh, S. *et al.* Comparison and validation of genomic predictors for anticancer drug sensitivity. *J. Am. Med. Inform. Assoc.* **20,** 597–602 (2013).

CHAPITRE 3

INTEGRATIVE PHARMACOGENOMICS TO INFER LARGE SCALE DRUG TAXONOMY

Pendant des décennies, le paradigme « un médicament — une cible — une maladie » a dicté une grande partie du processus de développement des médicaments. Cependant, au cours des dix dernières années, d'énormes progrès en génomique et transcriptomique ont graduellement changé cette vision simpliste du mécanisme d'action (MoA) vers un nouveau paradigme plus complexe, la pharmacologie des systèmes où un médicament peut se lier à plusieurs cibles. Plusieurs stratégies computationnelles ont été proposées pour élucider le mécanisme d'action des médicaments. Les approches traditionnelles prédisent les associations médicament-cible sur la base de la similarité chimique des médicaments ayant des cibles connues. Les approches bioinformatiques récentes ont générés des réseaux de similarité à partir de profils de transcription induits par les médicaments, ainsi de nouveaux mécanismes d'action pouvaient être inférées pour les médicaments dont le mécanisme est inconnu. Cependant, le groupement actuel des classes de médicaments est fondé sur des données pharmacologiques a priori, difficiles à recueillir pour de nouveaux composés. Il existe donc un besoin d'exploiter les nouvelles données pharmacogénomiques pour caractériser le MoA sans compter sur des informations difficiles à collecter, comme des indications thérapeutiques ou des effets secondaires. Dans notre étude, nous avons intégré différentes couches de données générées à partir d'un grand ensemble de données pharmacogénomiques afin de proposer de nouveaux MoA pour les composés chimiques. Ces couches de données sont (i) la similarité structurelle des médicaments (ii) les profils de perturbation transcriptomique induits par les médicaments à partir de la base de données LINCS; et (iii) les tests de viabilité cellulaire (sensibilité des lignées de cellules cancéreuses aux médicaments. Nous avons utilisé l'algorithme Similarity Network Fusion (SNF) pour intégrer efficacement ces trois couches et obtenir une couche fusionnée ou DNF. Nous avons constaté que la taxonomie des médicaments est améliorée significativement lorsque plusieurs caractéristiques liées au médicament sont fusionnées avec SNF. Nous avons classé correctement presque tous les inhibiteurs de kinases et suggéré de nouveaux mécanismes pour les autres composés non caractérisés. Notre approche innovante

met en évidence l'importance de l'intégration des données complémentaires concernant les médicaments tels que la chimie, la transcriptomique et les tests de viabilité cellulaire. DNF est générique et peut être facilement étendu à d'autres composés testes au laboratoire, en tant que tel, il constitue une ressource précieuse pour la communauté scientifique, spécifiquement la recherche de médicaments anticancéreux en fournissant de nouvelles hypothèses sur le MoA et une possibilité de repositionnement des médicaments. Notre outil combine toutes sortes de médicaments (en plus des médicaments antitumoraux). La disponibilité des essais biochimiques/cible criblée pour chaque médicament est certainement une limitation, mais qui sera levée dans le futur avec l'accroissement de la quantité de données pertinentes.

Contributions par auteur :

NEH et DMG sont responsable de la conception du projet, la collecte des données, l'analyse statistique, le code et l'implémentation du modèle en langage R, la rédaction du manuscrit, et l'interprétation des résultats. LSG a contribué au code en langage R, de la comparaison avec des méthodes concurrentes et de la reproductibilité de l'étude. ZS et PS ont contribué à la génération des données de transcriptomique à partir de la base LINCS L1000. RI et GB ont participé à la révision du manuscrit. AG et BHK ont supervisé l'étude, ont fourni les méthodes statistiques adéquates et ont participé à la révision du manuscrit. Tous les auteurs ont lu et approuvé le manuscrit final. Cet article est en révision dans *Nature communication.*

Nehme El-Hachem[1, 2,‡], Deena M.A. Gendoo[3, 4,‡], Laleh Soltan Ghoraie[3, 4], Zhaleh Safikhani[3, 4], Petr Smirnov[3], Ruth Isserlin[5], Gary D. Bader[5, 6,7], Anna Goldenberg[7, 8], Benjamin Haibe-Kains[3, 4,7, 8*]

[1] Integrative Computational Systems Biology, Institut de Recherches cliniques de Montréal, Montréal, Québec, Canada
[2] Department of Biomedical Sciences. Université de Montréal, Montréal, Québec, Canada
[3] Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada
[4] Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada
[5] The Donnelly Centre, Toronto, Ontario, Canada
[6]The Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada
[7] Department of Computer Science, University of Toronto, Toronto, Ontario, Canada
[8] Hospital for Sick Children, Toronto, Ontario, Canada
[‡] Co-first authors
[*] Corresponding author

3.1 Abstract

Identification of drug targets and mechanism of action (MoA) for new and uncharacterized drugs is important for optimization of drug efficacy. Current MoA prediction approaches largely rely on prior information including side effects, therapeutic indication and chemo-informatics. Such information is not transferable or applicable for newly identified small molecules. Accordingly, a systematic and unbiased approach towards MoA prediction is imperative to efficiently classify new compounds and infer their potential targets of MoA. We propose a method that only relies on basic drug characteristics, including drug structural information, drug perturbation and drug sensitivity profiles, which were not previously combined towards predicting drug targets and MoA. Using integrative computational pharmacogenomics, we implemented Drug Network Fusion (DNF), a scalable, integrative drug taxonomy. We demonstrate that DNF is effective towards prediction of drug targets and anatomical therapeutic chemical (ATC) classification, and enables robust inference of drug MoAs for new and existing compounds.

3.2 INTRODUCTION

Continuous growth and ongoing deployment of large-scale pharmacogenomic datasets has opened new avenues of research for the prediction of biochemical interactions of small drug molecules with their respective targets, also referred to as drug mechanisms of action (MoA). Several computational strategies have relied on chemical structure similarity to infer drug-target interactions[1–3], based on the assumption that structurally-similar drugs share similar targets, and ultimately, similar pharmacological and biological activity[4]. However, sole reliance on chemical structure information fails to consider drug-induced genomic and phenotypic perturbations, which directly connect with biological pathways and molecular disease mechanisms[5,6]. Recent approaches have thereby integrated drug-induced transcriptional profiles from Connectivity Map (CMAP)[7] into their algorithms, creating new ways for identification of drug-drug similarities and MoA solely based on gene expression profiles[8]. Other methods have integrated prior knowledge such as adverse effects annotations[9,10] and recent approaches showed that integrating multiple layers of

information had improved ATC prediction for FDA-approved drugs[11]. While these initiatives have undoubtedly paved great strides towards characterizing drug MoA, determining the consistency of such efforts towards prediction of new, uncharacterized small molecules remains a challenge.

The advent of high-throughput molecular profiling to identify patterns of small-molecule sensitivities across cell lines promises to shed additional insight into drug MoA. This type of drug bioactivity information remains largely unexploited in drug classification algorithms, despite its ongoing development over the past decade. The pioneering initiative of the NCI60 panel provided an assembly of tumour cell lines that have been treated against a diverse panel of over 100,000 small molecules[12,13]. The NCI60 dataset was the first large-scale resource enabling identification of lineage-selective small molecule sensitivities[14]. However, the relatively small number of 59 cancer cell lines of the NCI60 panel restricted the relevance of these data for prediction of drug MoA. The Cancer Therapeutics Response Portal (CTRP) has recently addressed this limitation by providing a resource of sensitivity measurements for extensively characterized cancer cell lines tested against a set of nearly 300 small molecules[15,16]. The latest CTRP release, coined CTRPv2, presents the largest quantitative *in vitro* sensitivity dataset available to date, spanning 860 cancer cell lines screened against a set of 481 small molecule compounds[16]. Individual assessment of these *in vitro* sensitivity datasets have highlighted their use towards determining mechanism of growth inhibition, and inference of MoA of compounds from natural products. It remains to be demonstrated, however, whether integration of these drug sensitivity data with other drug-related data, such as drug structures and drug-induced transcriptional signatures, can be used to systematically infer drug MoA.

Comprehensive molecular characterization of drug MoA for newly identified compounds requires high-throughput datasets that encapsulate a widespread range of drugs across multiple cancer cell lines. The aforementioned CTRPv2 sensitivity data perfectly qualifies for these requirements. However, such a dataset is unmatched by corresponding drug perturbation signatures from CMAP, which only characterizes 1309 drugs across 5 cancer cell lines. The CMAP project has recently been superseded by the L1000 dataset from the NIH Library of Integrated Network-based Cellular Signatures (LINCS) consortium[17], which has expanded upon the conceptual framework of CMAP and contains over 1.4 million gene

expression profiles spanning 20,413 chemical perturbations. Accordingly, the L1000 dataset provides an unprecedented compendium of both structural and transcriptomic drug data. A recent integrative study of the LINCS data showed that structural similarity are significantly associated with similar transcriptional changes, supporting the complementarity of these drug-related data[6].

To improve inference of drug MoA for new compounds, we leveraged our recent Similarity Network Fusion algorithm[18] to efficiently integrate drug structure, sensitivity, and perturbation data towards developing a large-scale molecular drug taxonomy, called Drug Network Fusion (DNF) **(Figure 3.1)**. DNF significantly outperformed taxonomies based on single data types at classifying drugs based on drug targets and therapeutic annotations. Our explorative analysis sheds light on how data integration approach can substantially improve characterization of MoA, both for existing drugs, but more specifically, for new compounds that lack deep pharmacological and biochemical characterization. Our results support DNF as a valuable resource to the cancer research community by providing new hypotheses on the compound MoA and potential insights for drug repurposing.

3.3 MATERIAL AND METHODS

A schematic overview of the analysis design is presented in **Figure 3.2**.

3.3.1 Processing of drug-related data and identification of drug similarity

Drug structure annotations : Canonical SMILES strings for the small molecules were extracted from PubChem [19], a database of more than 60 millions unique structures. Tanimoto similarity measures [20] between drugs were calculated by first parsing annotated SMILES strings for existing drugs through the *parse.smiles* function of the *rcdk* package (version 3.3.2). Extended connectivity fingerprints (hash-based fingerprints, default length 1,024) across all drugs was subsequently calculated using the *rcdk :: get.fingerprints* function[21].

Drug perturbation signatures : We obtained transcriptional profiles of cancer lines treated with drugs from the L1000 dataset recently released by the Broad Institute[22], which contains over 1.4 million gene expression profiles of 1000 ‘landmark’ genes across 20,413 drugs. We used our *PharmacoGx* package (version 1.1.4)[23] to compute signatures for the effect of drug

concentration on the transcriptional state of a cell, using a linear regression model adjusted for treatment duration, cell line identity, and batch to identify the genes whose expression is significantly perturbed by drug treatment:

$$G = \beta 0 + \beta iCi + \beta tT + \beta dD + \beta bB$$

where
G = *molecular feature expression (gene)*
Ci = *concentration of the compound applied*
T= *cell line identity*
D = *experiment duration*
B = *experimental batch*
βs = *regression coefficients.*

The strength of the feature response is quantified by βi. G and C are scaled variables (standard deviation equals to 1) to estimate standardized coefficients from the linear model. The transcriptional changes induced by drugs on cancer cell lines are subsequently referred to throughout the text as *drug perturbation signatures*. Similarity between estimated standardized coefficients of drug perturbation signatures was computed using the Pearson correlation coefficient, with the assumption that drugs similarly perturbing the same set of genes might have similar mechanisms of action.

Drug sensitivity signatures: We obtained summarized dose-response curves from the published drug sensitivity data of the NCI60 [14] and CTRPv2 [16] datasets integrated in the *PharmacoGx* package. We relied on the calculated Z-score and area under the curve (AUC) metrics for NCI60 and CTRPv2, respectively. Drug similarity was defined as the Pearson correlation of drug sensitivity profiles.

3.3.2 Development of a drug network fusion (DNF) taxonomy

We used our Similarity Network Fusion algorithm[18] to identify drugs that have similar mechanisms of actions by integrating three data types representing drug structure, drug perturbation, and drug sensitivity profiles. Drug structure and drug perturbation taxonomies were based on drug-drug similarity matrices computed from the PubChem SMILES and the the L1000 dataset, respectively. The drug sensitivity taxonomy was composed of the drug-drug similarity matrix of the sensitivity signatures extracted from either the NCI60 or CTRPv2 datasets. For each dataset, an affinity matrix was first calculated using the *affinityMatrix* function as described in the *SNFtool* package (version 2.2), using default parameters. We

combined the three affinity matrices of the structure, perturbation, and sensitivity taxonomies into a Drug Network Fusion (DNF) matrix using the *SNFtool::SNF* function **(Figure 3.2)**. Two separate DNF matrices were generated dependant on the sensitivity layer used (either CTRPv2 or NCI60). The developed DNF taxonomies, as well as the single data type taxonomies, were subsequently tested against benchmark datasets to validate their drug mode of action (MoA).

3.3.3 Assessment of drug mode of action across drug taxonomies

Drug-target associations. Known target associations for drugs pertaining to the NCI-60 dataset were downloaded from CHEMBL (file version 15-3-46-00) [24]. Drug-target associations for drugs of the CTRPv2 dataset were obtained from the CTRPv2 website (http://www.broadinstitute.org/ctrp.v2/?page=#ctd2Target). Drugs with annotated targets were filtered to retain only targets with at least two drugs.

Anatomical therapeutic classification system (ATC). ATC annotations[25] for the drugs common to the NCI60 and CTRPv2 datasets were downloaded from ChEMBL (file version 15-3-18-59)[24]. These ATC codes were filtered to retain only those categories with at least one pair of drugs sharing a pharmacological indication. The drugs with known ATC annotations from the NCI60 and CTRPv2 datasets were subsequently used as a validation benchmark against singular drug taxonomies and the DNF taxonomy.

Evaluation of drug mechanism of action across taxonomies

We assessed the predictive value of our developed taxonomies against drug-target and ATC benchmark datasets to determine the extent to which single data type taxonomies and the DNF taxonomy recapitulate known drug MoA **(Figure 3.3)**. We adapted the method from Cheng et al [26] to compare benchmarked datasets against singular drug taxonomies (Drug Perturbation, Drug Structure, or Drug Sensitivity) as well as the integrated DNF taxonomy. This method is further detailed below for the benchmark datasets used in our study. First, we created adjacency matrices that indicate whether each pair of drugs share a target molecule or ATC annotation. The drug-target and ATC adjacency matrices were then converted into a vector of similarities between every possible pair of drugs where the value '1' was assigned in the

vector if the paired drugs were observed the same target/ATC set, and '0' otherwise. Similarly, the affinity matrices of singular drug taxonomies as well as the DNF taxonomy matrix were converted into vectors of drug pairs, with the similarity value of the drug pairs retained from their original corresponding matrix. Binary vectors of the benchmarks were compared to the four continuous vectors of the drug taxonomies by computing the receiver-operating curves (ROC) using the *ROCR* package (version 1.0.7)[27], and the area under the curve (AUC) using the *concordance.index* function of the *survcomp* package (version 1.18.0)[28]. The AUC estimates the probability that, for two pairs of drugs, drugs that are part of the same drug set (same therapeutic targets or ATC functional annotations) have higher similarity than drugs that do not belong to the same drug set. AUC calculations for each of the four taxonomies were statistically compared against each other using the *survcomp::compare.cindex* function.

3.3.4 Detection of drug communities and visualization

Clusters of drug communities were determined from the DNF taxonomy using the affinity propagation algorithm[29,30] from the *apcluster* package (version 1.4.2). The apcluster algorithm generates non-redundant drug communities, with each community represented by an exemplar drug. An elevated q value parameter, which determines the quantiles of similarities to be used as input preference to generate small or large number of clusters, was set at q=0.9 within the *apcluster* function to produce a large number of communities. Networks of exemplar drugs were rendered in *Cytoscape* (version 3.3.0)[31]. Drug structures were rendered using the *chemViz* plugin (version 1.0.3) for cytoscape[32]. A minimal spanning tree of the exemplar drugs was determined using Kruskal's algorithm as part of the *CySpanningTree* plugin (version 1.1)[33] for cytoscape.

3.3.5 Research replicability

All the code and data links required to reproduce this analysis is publicly available on github.com/bhklab/DNF. The procedure to setup the software environment and run our analysis pipeline is also provided. This work complies with the guidelines proposed by Robert Gentleman[34] in terms of code availability and replicability of results.

3.4 RESULTS

We developed a large-scale molecular taxonomy, Drug Network Fusion (DNF), by integrating drug structure, drug sensitivity, and perturbation signatures using our recently developed Similarity Network Fusion algorithm[18] (Figure 3.2). Drug structure (SMILES representations) were extracted from the PubChem database, containing 60 million compounds. Drug perturbation signatures, representing drug-induced gene expression changes, were extracted from the recent LINCS L1000 dataset. Drug sensitivity signatures representing cell line viability across cancer cell lines were extracted from the CTRP portal, which contains pharmacological profiles of several hundred cell lines **(**Fig. III.I). We have tested the robustness of our approach by also generating a DNF taxonomy using the NCI60 sensitivity dataset, which contains pharmacological profiles for only 60 cell lines but spans thousands of drugs (Fig. III.I). Collectively, both tests serve to span a large spectrum of sensitivity signatures across both drug compounds and cancer cell lines. Using CTRPv2, our DNF taxonomy is composed of 239 drugs for which all of drug structure, drug perturbation, and drug sensitivity information could be fused. Using NCI60, a total of 238 common drugs were used. Notably, the overlap between the drugs of the NCI60 and CTRP datasets is small **(**64 drugs; Figure III. I**)** , which underscores the complementarity of these two datasets.

To demonstrate the benefit of our integrative approach, we assessed the predictive value of the DNF taxonomy for drug targets and functional classification and compared it to only using structure, sensitivity, or perturbation data alone. In addition, we used affinity propagation clustering (APC) on the DNF taxonomy to determine communities of drugs which share a similar MoA.

Performance of drug taxonomies against known drug targets

Determining drug-target interactions is important in the drug development process. The identification of new targets opens new avenues for drug repurposing efforts, and suggests new pathways and mechanisms by which drugs can operate in cells. Drug targets were identified for 193 drugs in CTRPv2 and these drugs were filtered to retain target categories with more than one drug, resulting in a set of 141 drugs available for benchmarking. Similarly, drug targets were identified for 101 drugs in NCI60, from which 73 drugs shared a target with at least another drug.

We assessed the predictive value of our single-data layer and integrative drug taxonomies against drug targets and ATC functional classification. We performed a ROC analysis to quantify how well our drug taxonomies align with established drug target (Figure 3.3). By statistically comparing the resulting AUC values, we were able to determine whether our integrative drug taxonomy outperformed taxonomies based on a singular data analyses (Figure 3.3). We tested how our integrated taxonomy using the CTRPv2 drug sensitivity taxonomy compares against single-layers for drug-target designations from ChEMBL (Figure 3.4A). Of the three single-layer taxonomies validated against annotated drug targets from CTRPv2, the drug sensitivity layer outperformed the structure and perturbation taxonomies (AUC of 0.83, 0.71 and 0.64 for sensitivity, structural and perturbation data layers, respectively) (Figure 3.4A). Importantly, DNF yielded the best predictive value (AUC of 0.89, Figure 3.4A), and was significantly higher than any single-layer taxonomy (one-sided t test p-value < 1E-16).

We replicated our integrative taxonomy approach using the set of drug sensitivity signatures obtained from the NCI60 dataset where a much smaller panel of cell lines has been screened (60 vs. 860 cell lines for NCI60 and CTRPv2, respectively). This integrative taxonomy (Figure III.II) was generated and validated against the drug-target benchmark from ChEMBL databases since no drug-targets annotation were provided from the NCI60 site. Our evaluation of single-layer taxonomies demonstrates that drug similarities based on sensitivity signatures were the most efficient in predicting drug-target associations (AUC of 0.69; Figure III.IIA) compared to structure and perturbation (AUC of 0.61 and 0.49, respectively; Figure III.IIA). DNF was significantly more predictive of drug-target associations compared to single-layer taxonomies from structure and perturbation but not sensitivity (AUC of 0.70 and one-sided superiority test p-values < 0.05, Figure III.IIA).

Performance of drug taxonomies against known functional classes

Predicting the anatomical classification (ATC) of a drug provides existing and new insights about its pharmacological mechanism, and ultimately presents new potential indications for previously uncharacterized drugs. ATC codes were identified for 59 and 122 drugs pertaining to CTRPv2 and NCI60, respectively. These codes were filtered to retain only those categories with at least one pair of drugs sharing a pharmacological indication. A total of 43 and 88 drugs

with known ATC annotations from the CTRPv2 and NCI60 datasets, respectively, were subsequently used for performance assessment.

We conducted a second validation of our taxonomies against ATC drug classification (Figure 3.4B). Drug sensitivity was not the most predictive layer for ATC classification and exhibited comparable predictive power as drug perturbation (Figure 3.4B). The structure-based taxonomy (Figure 3.4B) was the most predictive amongst single-layer taxonomies (AUC of 0.72, 0.57 and 0.54 for structure, sensitivity, and perturbation layers, respectively). The integrative drug taxonomy significantly outperformed single-layer taxonomies (AUC of 0.77 with one-sided t test p-value < 0.05; Figure 3.4B). Interestingly, DNF outperforms single-layer taxonomies when tested for functional classification based on ATC (AUC of 0.87 with one-sided t test p-values < 0.05; Figure III.IIB). Similarly, as observed with CTRPv2, structural similarity remains the best performing single-layer taxonomy when tested against ATC classification. .

Identification of Drug Communities Using DNF Taxonomy

To assess the biological relevance of integrative drug taxonomy in discovering drugs with similar MoA, we applied the affinity cluster propagation algorithm[30] to identify clusters of highly similar drugs referred to as *drug communities* (Figure 3.5, Figure III.III). These communities can be represented by their most representative drug and the similarities between communities represented a network where each node is labeled by the exemplar drug. Our initial analysis of the DNF taxonomy based on CTRPv2 sensitivity identified 53 communities. Of these, we identified 39 drug communities, which have at least two drugs with a known mechanism of action.

Overall, our integrative taxonomy developed using the CTRPv2 has produced a substantial and consistent classification of drugs for a variety of functional classes (Table III.I). Briefly our classifications recapitulate most of the protein target-drug associations represented in CTRPv2: Receptor tyrosine kinases and non-receptor tyrosine kinases (including EGFR, VEGFR, ALK, ABL1, SRC, RAF, MEK, IGFR-1) inhibitors, PI3K/mTOR family inhibitors, proapoptotic (including the p53 tumor suppressor) and anti-apoptotic (including the MDM2 and BCL-2 oncogenes) inhibitors, epigenetic regulators (HDACs) inhibitors, glycosyltransferase NAMPT inhibitors, cell cycle kinases inhibitors (CDKs, PLK,

ATM), DNA replication (topoisomerases), repair and synthesis (TYMS) inhibitors, HMG CoA and proteasome inhibitors.

We replicated our integrative taxonomy using the NCI60 sensitivity dataset (Figure III.I), and identified 51 communities, of which 20 communities showed at least two drugs with a known mechanism of action. We are aware that an important number of drugs has unannotated target and ATC codes, as most of the drugs in this study are experimental or uncharacterized chemicals in NCI60, however for reproducibility and validation concerns we did not manually annotated our compound collections.

3.5 DISCUSSION

Identification of MoA for newly uncharacterized compounds is a key challenge towards characterizing on-targets responsible of pharmacological effect and off-targets associated with unexpected physiological effects. Shortcomings of current approaches include a degree of reliance on pharmacological, biochemical, and functional annotations that pertain to existing, well-characterized drugs, and which may not be applicable towards prediction of a new small compounds (Figure 3.6)[11,35]. Compounding this issue is the absence of a high-throughput, integrative classification that merges basic complementary drug characteristics, such as chemical structure, *in vitro* drug sensitivity and transcriptional perturbation signatures. In addition to hindering efficient classification of new, uncharacterized drugs, such shortcomings also pose an obstacle towards proper evaluation of the current methods for drug taxonomy inference. In particular, external data cannot be used to simultaneously develop the method and independently assess its quality. Our analysis addresses these issues by conducting, to our knowledge, the first large-scale integration of drug structure, sensitivity and perturbation signatures towards prediction of drug MoA.

We capitalized upon our integrative Similarity Network Fusion method[18] to construct a high-throughput drug similarity network (DNF), based on the fusion of drug structure, sensitivity, and perturbation data. The construction of drug-similarity networks and their subsequent fusion allows us to fully harness the complementary nature of several drug datasets, and generate an informative clustering of drugs across multiple data types. Testing how well different drug taxonomies correctly predict drug targets (Figure 3.4A) and

anatomical (ATC) drug classifications (Figure 3.4B) indicates that DNF constitutes a marked improvement towards drug classification, compared to single data type analyses using either drug sensitivity, structure, or perturbation information alone. This observation is sustained even with the use of a different source of *in vitro* sensitivity data (Figure III.I) to generate the DNF matrix. Accordingly, our integrative approach succeeds in combining several drug data types into a single comprehensive network that represents the full spectrum of the underlying data.

Relying on drug-related data that only encompasses drug sensitivity, structure, and perturbation profiles ultimately presents a flexible approach towards comprehensive drug classification (Figure 3.6). We have removed any reliance on existing pharmacological, biochemical, or functional annotations that pertain to existing drugs, such as drug-target classifications or knowledge of the anatomical and organ system targeted by the drug compounds. Accordingly, our DNF method only requires basic drug information, including drug structures, sensitivity, and perturbation profiles, to determine drug MoA. These types of data, compared to other mechanistic annotations including ATC or drug target information, are much easier to generate for newly uncharacterized compounds, which ultimately facilitates proper characterization of new compounds. This provided a more extensive characterization of the compound across multiple manifolds of drug associations, and ultimately allowed us to test our DNF drug associations against both drug-target and anatomical therapeutic classifications (ATC).

Comparing our integrative DNF taxonomy with single data layers revealed the importance of drug sensitivity information towards improving prediction performance of drug-target associations (Figure 3.4A). Such findings support the relevance of bioactivity assays to predict drug targets, and underscore the comprehensive nature of the CTRPv2 dataset (860 cell lines screened with 16 drug concentrations, tested in duplicate)[16]**.** Similarity, we have observed a priority for drug structure information towards prediction of ATC drug classification (Figure 3.4B). Our approach thus exemplifies how DNF and singular taxonomies are compared against a number of drug benchmarks, and highlights the interplay between different types of data for generating relevant drug classifications.

The DNF taxonomy highlights many cases of drug clusters with known MoA, capturing context-specific features associated to drug sensitivity and genomic profiles in

cancer cells (Figure 3.5). These cases, to some extent, serve as experimental validation of our method. We classified correctly all BRAF (V600E mutation) inhibitors, which include drugs already tested in metastatic melanoma (community C18: dabrafenib, GDC0879, PLX4720; (Figure 3.5) and mitogen-activated protein kinase/ERK kinase (MEK) inhibitors (C41: namely trametinib and selumetinib; Figure 3.5). BRAF regulates the highly conserved MAPK/ERK signaling pathway, and BRAF mutational status has been proposed as a biomarker of sensitivity towards selumetinib and other MEK inhibitors[36,37]. This explains the tight connection of these two communities (Figure 3.5).

The DNF taxonomy also represents a new and comprehensive resource that can be mined to uncover new relationships between drugs and mechanisms of action. We identified a community of HMG Co-A reductase inhibitors (statins) composed of fluvastatin, lovastatin, and simvastatin (C30; Figure 3.5). These are a class of cholesterol-lowering drugs, and which have been found to reduce cardiovascular disease. Interestingly, parthenolide clusters with this community, and has been experimentally observed to inhibit the NF-Kb inflammatory pathway in atherosclerosis and in colon cancer[38,39], thereby suggesting similar behavior to statin compounds. We also classified correctly drugs with unannotated mechanisms/targets in CTRPv2 such as ifosfamide, cyclophosphamide and procarbazine (C17; Figure 3.5) which are known alkylating agents (ATC code : L01A). Furthermore, this was also true for docetaxel and paclitaxel (C21; Figure 3.5), two taxanes drugs with unannotated target in CTRPv2 although known as sharing similar MoA (ATC code : L01CD).

Our integrative drug taxonomy is also able to identify targets for drugs with poorly understood mechanisms and to infer new mechanism for other drugs. Community C15, for example, contains tigecycline and Col-3 (Figure 3.5); both are derivatives of the antibiotic tetracycline[40]. Tigecycline is an approved drug, however its target is not characterized in humans. Col-3 showed antitumor activities by inhibiting matrix metalloproteinase[40]. Interestingly, tosedostat (CHR-2797), a metalloenzyme inhibitor with antiproliferative potential, is also a member of this community[41]. Another drug in this community, phloretin, is a natural compound with uncharacterized targets and has been recently shown to deregulate matrix metalloproteinases at both gene and protein levels[42]. Our results suggest that matrix metalloproteinases would be the preferred target for drugs in this community, supporting the need for further experimental investigation. DNF also consolidated previous findings for drugs

that may serve as tubulin polymerization disruptors, and which have not been previously classified as such. We identified a community of three drugs (C49) in which LY2183240, and YK-4-279 have been recently identified to decrease alpha-tubulin levels[16]. Tivantinib, a c-MET tyrosine kinase inhibitor, also blocked microtubule polymerization[43]. Interestingly, this community is tightly connected to known microtubule perturbators (community C21; Figure 3.5).

Our results also concur with the study of Rees et al.[44] regarding cluster of the BCL-2 inhibitors ABT-737 and navitoclax (community C33; Figure 3.5), where the authors reported that a high expression of BCL-2 confers sensitivity to these two drugs. This was not the case for another BCL-2 inhibitor, obatoclax. They proposed that a metabolic modification of obatoclax in cells impacts its interaction with BCL-2 proteins, therefore reducing its potency. We showed indeed that obatoclax did not cluster with the other two BCL-2 inhibitors (ABT-737 and navitoclax). Such an example demonstrates how the structural and sensitivity profiles of these two BCL-2 inhibitors are largely coherent in contrast to obatoclax, which previously showed off-target effects compared to ABT-737[45]. This provides a good evidence to consider sensitivity profiles when developing new potent and specific BCL-2 inhibitors.

Our results suggest the existence of "super communities", that are a grouping of several communities contributing to a larger, systems-based MoA. An example is provided by the tightly connected communities C3, C21, C23, C43 (Figure 3.5). One of these communities (C3: Alvocidib, PHA-793887 and staurosporine) includes well-characterized inhibitors of cyclin dependant kinases (CDKs) that are known to be major regulators of the cell cycle. BMS-345541 for example, which also clusters with drugs in C3, is an ATP non-competitive allosteric inhibitor of CDK[46]. Those compounds are positioned close in the community network to topoisomerase I and II inhibitors (C43: SN-38, topotecan, etoposide, teniposide), microtubule dynamics perturbators (C21: paclitaxel, docetaxel, vincristine, parbendazole) and polo-like kinase inhibitors (C23 : GSK461364, GW843682X). Iorio *et al.*, reported that the similarity between CDK inhibitors and the other DNA-damaging agents is mediated through a p21 induction, which explains the interconnection and rationale of similar transcriptional and sensitivity effects of these regulators of cell cycle progression[8].

Our results recapitulate findings from clinical trials. For example, Ibrutinib, which is a Bruton tyrosine kinase inhibitor (BTK) approved for the treatment of Mantle cell lymphoma

and chronic lymphocytic leukemia, clustered with the known EGFR inhibitors (C2: erlotinib, gefitinib, afatinib and others). The effect of ibrutinib in EGFR Mutant Non-Small Cell Lung Cancer has been reported in a recent clinical trial (ClinicalTrials.gov Identifier : NCT02321540). This was also the case for MGCD265, a Met inhibitor, which clustered with most of the VEGFR (vascular endothelial growth factor receptor) inhibitors (C50: pazopanib, cediranib and others). In this community, pazopanib is the only FDA approved drug for the treatment of renal cell carcinoma. There exist a recent evidence that the clinical drug candidate MGCD265 has an application in many types of cancers including renal cell carcinoma (ClinicalTrials.gov Identifier : NCT00697632).

Our study suggests that drug sensitivity data is an important asset for computational methods that predict drug mechanism of action. To test the robustness of the fusion algorithm with respect to the scale of the drug sensitivity profiles, we also applied our methodology on the NCI60 dataset, which comprises a much smaller panel of cell lines (60 vs. 860 for NCI60 and CTRPv2, respectively). The NCI60 panel compensates for its small cell line panel by the large number of screened drugs (>40,000 drugs tested on the full panel; Figure III. I). Testing DNF using the NCI60 sensitivity information reveals that our integrative taxonomy continues to supersede single-layer drug taxonomies across the target and ATC benchmarks (Figure III.II). Interestingly, some of the identified communities using NCI60, such as the tight connection between BRAF/MEK inhibitor drugs (C42; Figure III.III), had also been identified in our original analysis using CTRPv2 sensitivity profiles. This demonstrates a high degree of specificity of drug-target associations across cell lines and experimental platforms, which is crucial in biomarker identification and translational research.

The DNF taxonomy encompassing the NCI60 dataset has also identified a number of well-characterized drug communities. These include the community composed of EGFR inhibitors (C20; Figure III.III). Our results for community C14 (cardiac glycosides) also concur with the study of Khan et al [5] (Figure III.III). These compounds inhibit Na+/K+ pumps in cells. Using a 3D chemical descriptor approach combined with genomic features, Khan *et al* had also identified bisacodyl, a laxative drug, as sharing a similar mechanism with cardiac glycosides, despite its structural dissimilarity to that class of compounds[5]. Notably, our integrative taxonomy recapitulates these findings, which demonstrates that combination of

structural and genomic drug information is a promising strategy towards elucidating drug mechanisms.

Our DNF based on NCI60 sensitivity information enabled identification of new drugs with uncharacterized MoA that we believe warrant further experimental investigation. We found that communities C2, C5, C32, and C51 were closely connected (Figure III.III). These communities contain a number of compounds, which showed antitumor activity by generating reactive oxygen species (e.g. C2: elesclomol, fenretinide; C5 : ethacrynic acid, curcumin; C32 : bortezomib, menadione; C51 : celastrol, withaferin A, parthenolide, thapsigargin). Interestingly, ethacrynic acid, an FDA approved drug indicated for hypertension, clustered with curcumin, a component of turmeric. Ethacrynic acid inhibits glutathione S-transferase (GSTP1) and induced mitochondrial dependant apoptosis through generation of reactive oxygen species (ROS) and induction of caspases[47]. Curcumin showed antitumor activity by production of ROS and promotion of apoptotic signaling. Thus, we suggest that GSTP1 could be a potential target of the widely used natural compound curcumin.

In conclusion, we have developed Drug Network Fusion (DNF), an integrative taxonomy inference approach leveraging the largest quantitative compendiums of structural information, pharmacological phenotypes and transcriptional perturbation profiles to date. We used DNF to conduct a cross-comparative assessment between our integrative taxonomy, and single-layer drug taxonomies based on drug structure, perturbation, or sensitivity signatures. Our exploratory analysis indicates the superiority of DNF towards drug classification, and also highlights singular data types that are pivotal towards prediction of drug categories in terms of anatomical classification as well as drug-target relationships. Overall, the DNF taxonomy has produced a consistent classification of drugs for multiple functional classes in both CTRPv2 and NCI60 (Table III.I). The comprehensive picture of drug-drug relationships produced by DNF has also succeeded in identifying new and potentially interesting drug MoA. The integrative DNF taxonomy has the potential to serve as a solid framework for future studies involving inference of MoA of new, uncharacterized compounds, which represents a major challenge in drug development for precision medicine.

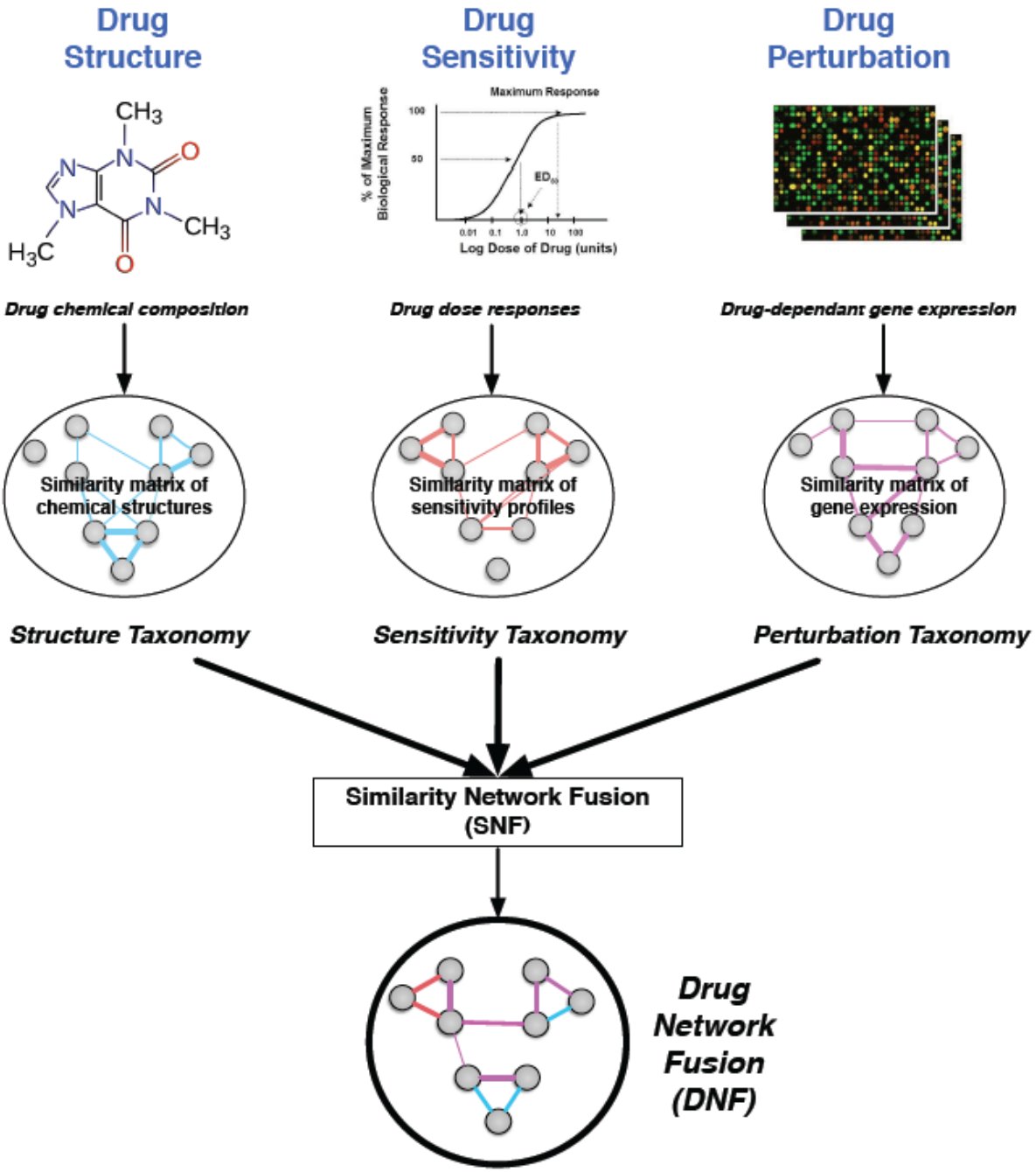

Figure 3.1 : Schematic representation of the SNF method and its use towards integration of different types of drug information. Datasets representing drug similarity, drug sensitivity, and drug perturbation profiles are first converted into drug-drug similarity matrices. Similarity matrices are fully integrated within the SNF method to generate a large-scale,multi-tier, Drug Fusion Network (DNF) taxonomy of drug-drug relationships.

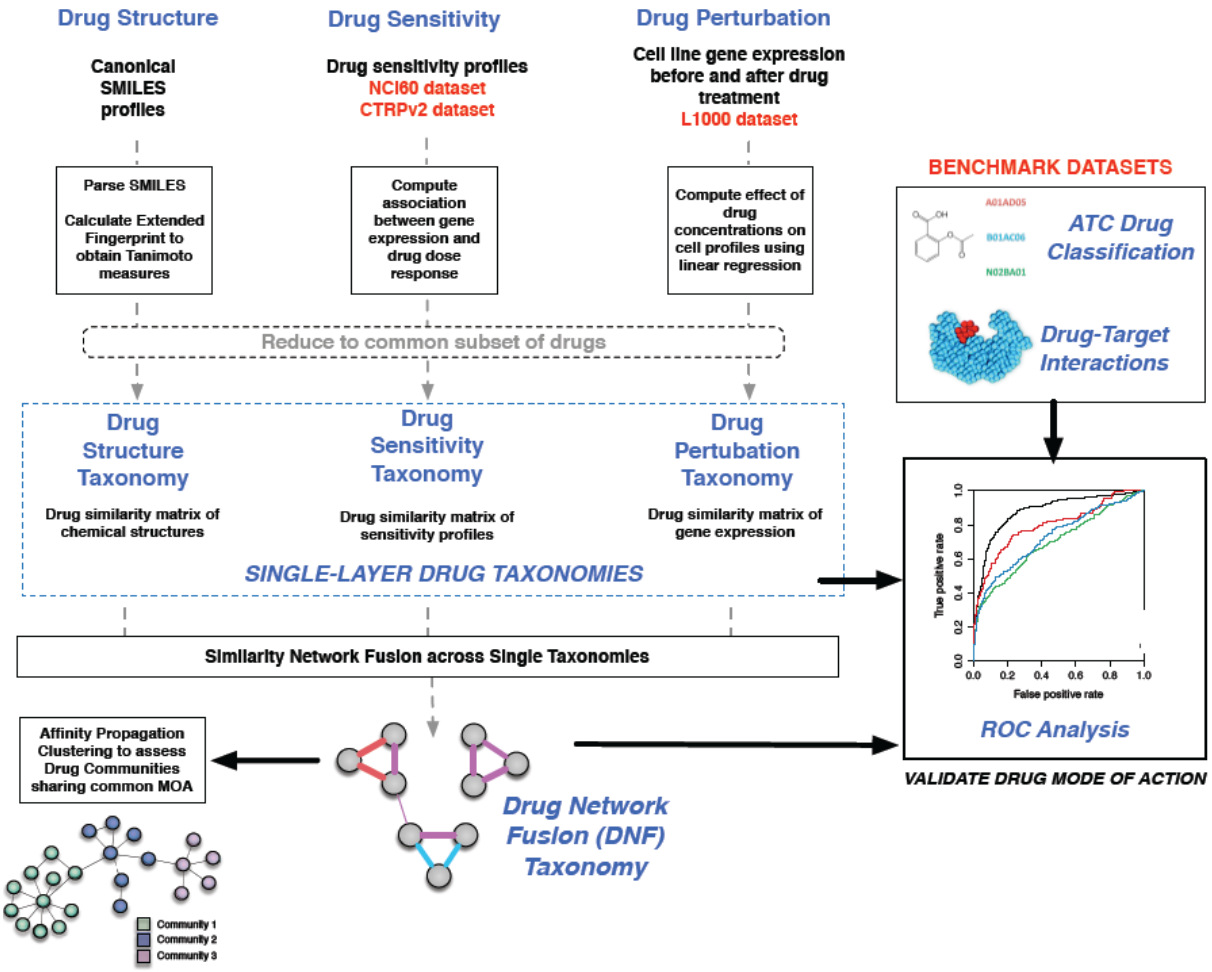

Figure 3.2 : Overview of the study design. Drug sensitivity profiles from the NCI60 and the CTRPv2 datasets, along with drug perturbation and drug structure data from the L1000 dataset, are first parsed into drug-drug similarity matrices that represent single-dataset drug taxonomies. Two DNF taxonomies are generated using the drug perturbation and drug structure taxonomies and the drug sensitivity taxonomy from either the NCI60 or CTRPv2 datasets. DNF taxonomies and singledataset taxonomies are tested against benchmarked datasets containing ATC drug classification and drug-target information, to validate their efficacy in predicting drug MoA. Additional clustering is conducted on DNF taxonomies to identify drug communities sharing a MoA.

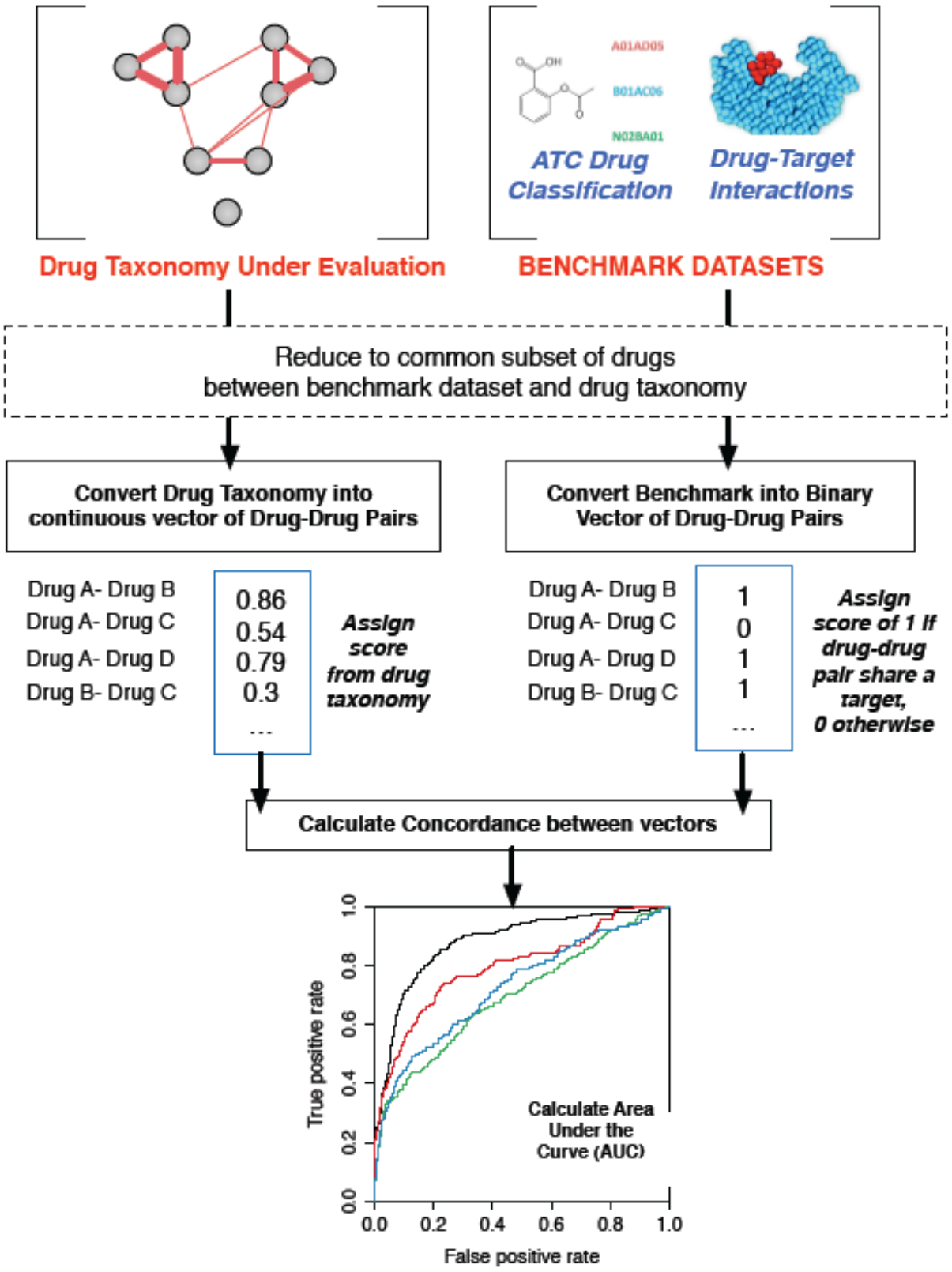

Figure 3.3 : Schematic representation of the validation of the DNF and single data type

analyses against drug benchmarks. Drug taxonomies are converted into a continuous vector of drug-drug pairs. Benchmark datasets are converted into binary vectors, whereby a given drug-drug pair is assigned a value of '1' if the drugs share a common drug target or ATC classification, and '0' otherwise. Vectors are compared using the concordance index and the area under the curve (AUC) is calculated from the receiver-operating curves (ROCs).

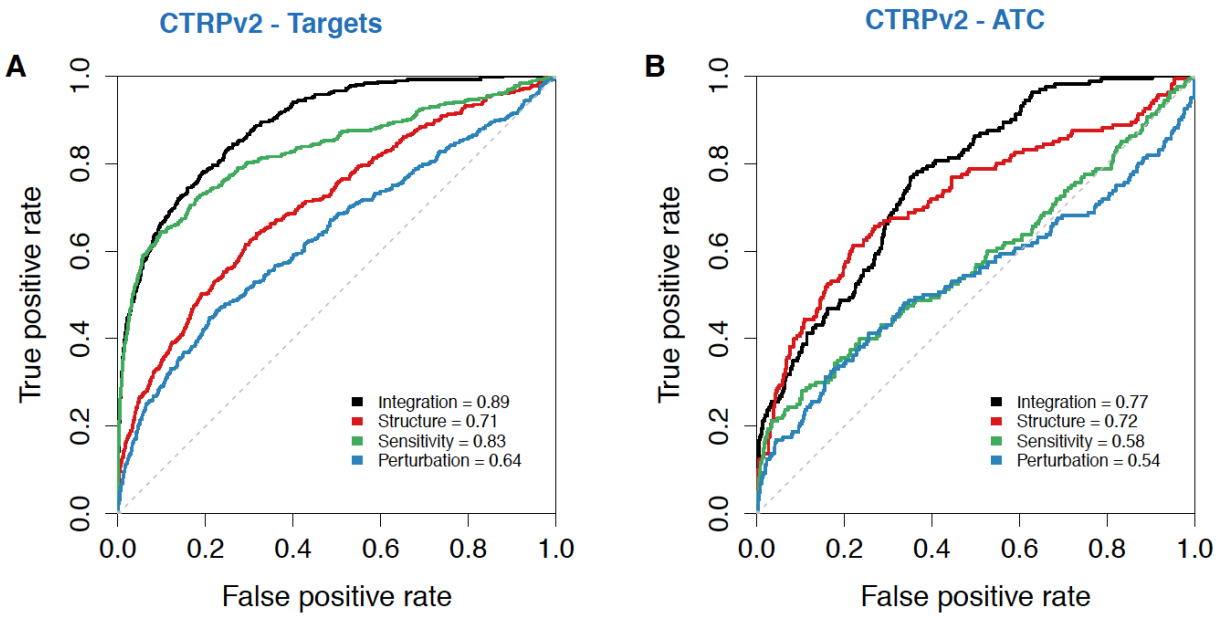

Figure 3.4 : Validation of the DNF taxonomy (using CTRPv2 sensitivity data) and single dataset taxonomies against the ATC and Drug-target benchmarks. ROC curves are shown for each of the taxonomies generated with the CTRPv2 sensitivity dataset, tested against ATC annotations and drug-target information from Chembl or internal benchmarks. A diagonal (red) representing the null case (AUC=0.5) is drawn for clarity. A) ROC curve against drug-targets B) ROC curve against ATC drug classifications.

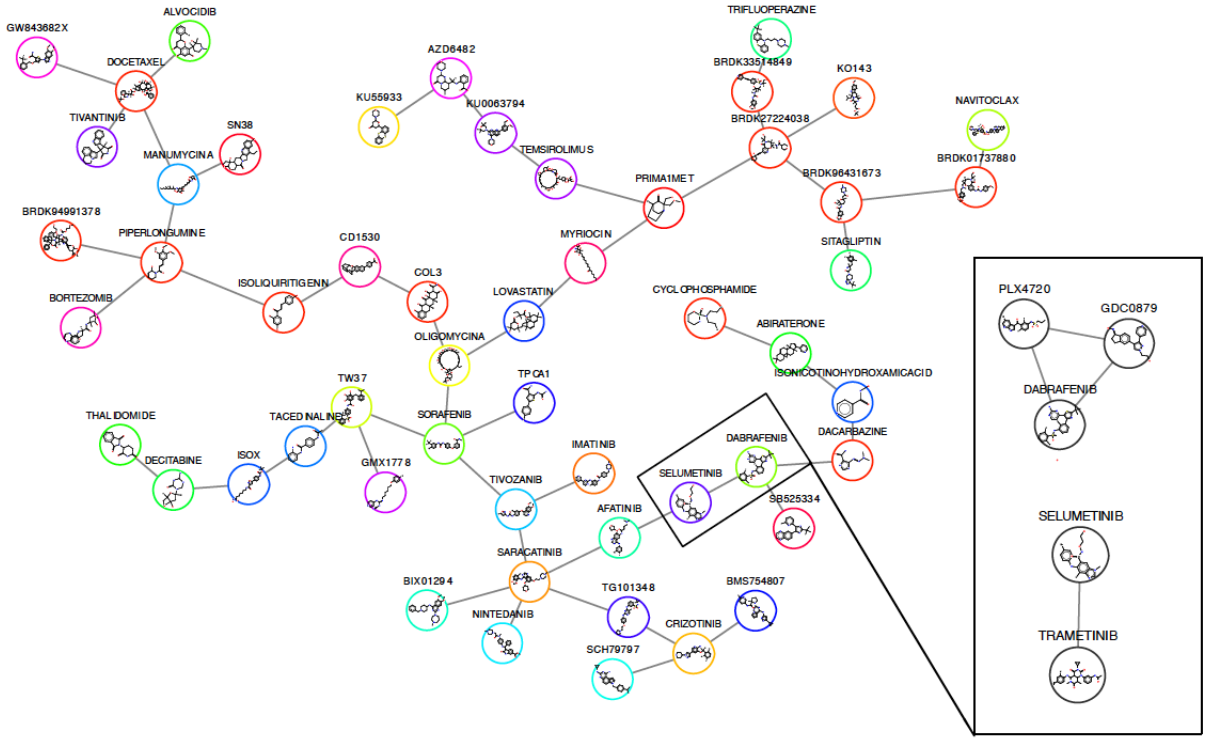

Figure 5 : Network representation of 51 exemplar drugs that are representative of the drug communities identified by the DNF taxonomy (using CTRPv2 sensitivity data). Each node represents the exemplar drugs, and node sizes reflect the size of the drug community represented by the exemplar node. Nodes are colored to reflect shared MoA as determined using the drug-target benchmark used for Figure 4. Communities sharing similar MoA and proximity in the network are highlighted, with the community number indicated next to each community. Drug communities pertaining to the super-community are labelled in red.

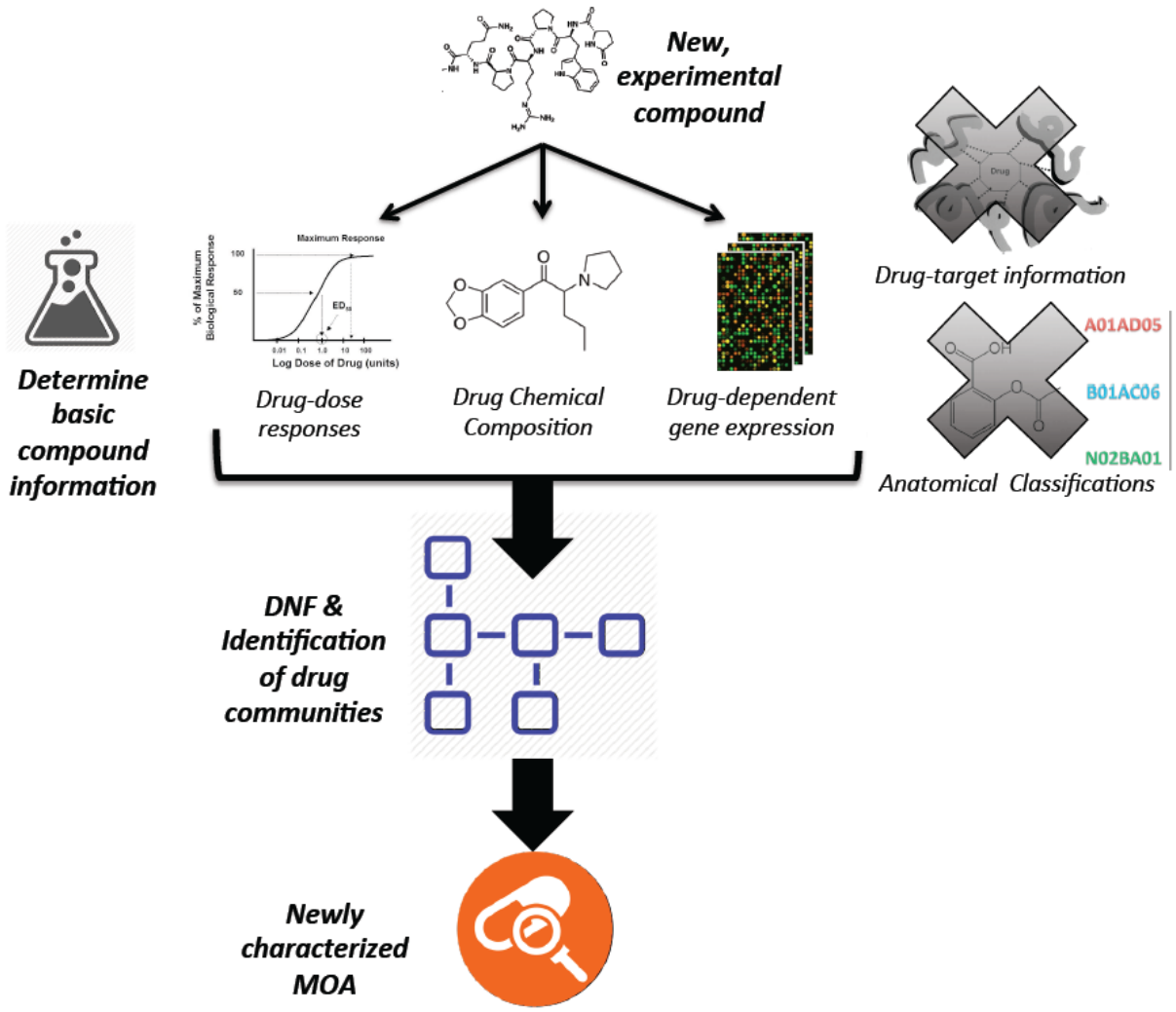

Figure 3.6 : Schematic of the adaptability of DNF towards prediction of new experimental compounds.

References

1. Sheridan RP, Kearsley SK. Why do we need so many chemical similarity search methods? Drug Discov Today. 2002; 7 : 903–911.

2. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, et al. Predicting new molecular targets for known drugs. Nature. 2009; 462 : 175–181.

3. Dunkel M, Günther S, Ahmed J, Wittig B, Preissner R. SuperPred: drug classification and target prediction. Nucleic Acids Res. 2008; 36 : W55–9.

4. Martin YC, Kofron JL, Traphagen LM. Do structurally similar molecules have similar biological activity? J Med Chem. 2002; 45 : 4350–4358.

5. Khan SA, Virtanen S, Kallioniemi OP, Wennerberg K, Poso A, Kaski S. Identification of structural features in chemicals associated with cancer drug response: a systematic data-driven analysis. Bioinformatics. 2014; 30 : i497–504.

6. Chen B, Greenside P, Paik H, Sirota M, Hadley D, Butte AJ. Relating Chemical Structure to Cellular Response: An Integrative Analysis of Gene Expression, Bioactivity, and Structural Data Across 11,000 Compounds. CPT Pharmacometrics Syst Pharmacol. 2015; 4 : 576–584.

7. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006; 313 : 1929–1935.

8. Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, Ferriero R, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. Proc Natl Acad Sci U S A. 2010;107 : 14 621–14 626.

9. Campillos M, Kuhn M, Gavin A-C, Jensen LJ, Bork P. Drug target identification using side-effect similarity. Science. 2008; 321 : 263–266.

10. Kuhn M, Al Banchaabouchi M, Campillos M, Jensen LJ, Gross C, Gavin A-C, et al. Systematic identification of proteins that elicit drug side effects. Mol Syst Biol. 2013; 9 : 663.

11. Napolitano F, Zhao Y, Moreira VM, Tagliaferri R, Kere J, D'Amato M, et al. Drug repositioning: a machine-learning approach through data integration. J Cheminform. 2013; 5 : 30.

12. Driscoll JS. The preclinical new drug research program of the National Cancer Institute. Cancer Treat Rep. 1984; 68 : 63–76.

13. Luna A, Rajapakse VN, Sousa FG, Gao J, Schultz N, Varma S, et al. rcellminer: exploring molecular profiles and drug response of the NCI-60 cell lines in R. Bioinformatics. 2015; doi:10.1093/bioinformatics/btv701

14. Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. Nat Rev Cancer. 2006; 6 : 813–823.

15. Basu A, Bodycombe NE, Cheah JH, Price EV, Liu K, Schaefer GI, et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. Cell. 2013; 154 : 1151–1161.

16. Seashore-Ludlow B, Rees MG, Cheah JH, Cokol M, Price EV, Coletti ME, et al. Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. Cancer Discov. 2015; 5 : 1210–1223.

17. NIH, Broad Institute. The LINCS Connectivity Map Project. In : The LINCS Connectivity Map Project [Internet]. 2015 [cited 2016]. Available : https://clue.io/

18. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014; 11 : 333–337.

19. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem Substance and Compound databases. Nucleic Acids Res. 2016; 44 : D1202–13.

20. Tanimoto TT. An Elementary Mathematical Theory of Classification and Prediction. International Business Machines Corporation; 1958.

21. Guha R, Others. Chemical informatics functionality in R. J Stat Softw. 2007; 18 : 1–16.

22. Duan Q, Flynn C, Niepel M, Hafner M, Muhlich JL, Fernandez NF, et al. LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. Nucleic Acids Res. 2014; 42 : W449–60.

23. Smirnov P, Safikhani Z, El-Hachem N, Wang D, She A, Olsen C, et al. PharmacoGx: An R package for analysis of large pharmacogenomic datasets. Bioinformatics. 2015; doi:10.1093/bioinformatics/btv723

24. Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, et al. The ChEMBL bioactivity database : an update. Nucleic Acids Res. 2014; 42 : D1083–90.

25. Nahler G. Anatomical therapeutic chemical classification system (ATC). In : Nahler G, editor. Dictionary of Pharmaceutical Medicine. Springer Vienna; 2009. pp. 8–8.

26. Cheng J, Xie Q, Kumar V, Hurle M, Freudenberg JM, Yang L, et al. Evaluation of analytical methods for connectivity map data. Pac Symp Biocomput. 2013; 5–16.

27. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005; 21 : 3940–3941.

28. Schröder MS, Culhane AC, Quackenbush J, Haibe-Kains B. survcomp : an R/Bioconductor package for performance assessment and comparison of survival models. Bioinformatics. 2011; 27 : 3206–3208.

29. Bodenhofer U, Kothmeier A, Hochreiter S. APCluster: an R package for affinity propagation clustering. Bioinformatics. 2011; 27 : 2463–2464.

30. Frey BJ, Dueck D. Clustering by passing messages between data points. Science. 2007; 315 : 972–976.

31. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003; 13 : 2498–2504.

32. chemViz2 : Cheminformatics App for Cytoscape [Internet]. [cited 1 Mar 2016]. Available : http://www.cgl.ucsf.edu/cytoscape/chemViz2/index.shtml

33. Shaik F, Bezawada S, Goveas N. CySpanningTree: Minimal Spanning Tree computation in Cytoscape. F1000Res. 2015; 4. doi:10.12688/f1000research.6797.1

34. Gentleman R, Lang DT. Statistical analyses and reproducible research. J Comput Graph Stat. Taylor & Francis; 2012; Available : http://amstat.tandfonline.com/doi/abs/10.1198/106186007X178663

35. Ma'ayan A, Rouillard AD, Clark NR, Wang Z, Duan Q, Kou Y. Lean Big Data integration in systems biology and systems pharmacology. Trends Pharmacol Sci. 2014; 35 : 450–460.

36. Friday BB, Yu C, Dy GK, Smith PD, Wang L, Thibodeau SN, et al. BRAF V600E disrupts AZD6244-induced abrogation of negative feedback pathways between extracellular signal-regulated kinase and Raf proteins. Cancer Res. 2008; 68 : 6145–6153.

37. Solit DB, Garraway LA, Pratilas CA, Sawai A, Getz G, Basso A, et al. BRAF mutation predicts sensitivity to MEK inhibition. Nature. 2006; 439 : 358–362.

38. Riganti C, Doublier S, Costamagna C, Aldieri E, Pescarmona G, Ghigo D, et al. Activation of nuclear factor-kappa B pathway by simvastatin and RhoA silencing increases doxorubicin cytotoxicity in human colon cancer HT29 cells. Mol Pharmacol. 2008; 74 : 476–484.

39. López-Franco O, Hernández-Vargas P, Ortiz-Muñoz G, Sanjuán G, Suzuki Y, Ortega L, et al. Parthenolide modulates the NF-kappaB-mediated inflammatory responses in experimental atherosclerosis. Arterioscler Thromb Vasc Biol. 2006; 26 : 1864–1870.

40. Syed S, Takimoto C, Hidalgo M, Rizzo J, Kuhn JG, Hammond LA, et al. A phase I and pharmacokinetic study of Col-3 (Metastat), an oral tetracycline derivative with potent matrix metalloproteinase and antitumor properties. Clin Cancer Res. 2004; 10 : 6512–6521.

41. Krige D, Needham LA, Bawden LJ, Flores N, Farmer H, Miles LEC, et al. CHR-2797 : an antiproliferative aminopeptidase inhibitor that leads to amino acid deprivation in human leukemic cells. Cancer Res. 2008; 68 : 6669–6679.

42. Ma L, Wang R, Nan Y, Li W, Wang Q, Jin F. Phloretin exhibits an anticancer effect and enhances the anticancer ability of cisplatin on non-small cell lung cancer cell lines by regulating expression of apoptotic pathways and matrix metalloproteinases. Int J Oncol. 2016; 48 : 843–853.

43. Katayama R, Aoyama A, Yamori T, Qi J, Oh-hara T, Song Y, et al. Cytotoxic activity of tivantinib (ARQ 197) is not due solely to c-MET inhibition. Cancer Res. 2013; 73 : 3087–3096.

44. Rees MG, Seashore-Ludlow B, Cheah JH, Adams DJ, Price EV, Gill S, et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. Nat Chem Biol. 2016; 12 : 109–116.

45. Vogler M, Weber K, Dinsdale D, Schmitz I, Schulze-Osthoff K, Dyer MJS, et al. Different forms of cell death induced by putative BCL2 inhibitors. Cell Death Differ. 2009; 16 : 1030–1039.

46. Bogoyevitch MA, Fairlie DP. A new paradigm for protein kinase inhibition: blocking phosphorylation without directly targeting ATP binding. Drug Discov Today. 2007; 12 : 622–633.

47. Wang R, Liu C, Xia L, Zhao G, Gabrilove J, Waxman S, et al. Ethacrynic Acid and a Derivative Enhance Apoptosis in Arsenic Trioxide--Treated Myeloid Leukemia and Lymphoma Cells: The Role of Glutathione S-Transferase P1-1. Clin Cancer Res. AACR; 2012; 18 : 6690–6701.

CHAPITRE 4

Characterization of conserved toxicogenomic responses in chemically exposed hepatocytes across species and platforms

Les études de micropuces à grande échelle sont utilisées pour identifier les changements transcriptionnels induits par les médicaments et les facteurs chimiques environnementaux. Dans ce contexte, le projet TG-GATES a généré des données de micropuces à partir d'échantillons de foie de rat et hépatocytes primaires (rat et humain) exposés à plus de 100 produits chimiques différents.

Le principal objectif de ce chapitre est d'évaluer la capacité des modèles cellulaire à récapituler les changements induits par les produits chimiques in vivo. Nous avons analysé le jeu de données TG-gates pour comparer les réponses transcriptionnelles précoces observées dans le foie des rats traités avec un grand nombre de produits chimiques à ceux provoqués avec les mêmes composés in vitro.

Nous avons développé un nouveau pipeline d'analyse qui combine efficacement l'analyse d'enrichissement (GSEA) en utilisant les voies de signalisation à partir de la base de données Reactome et une technique de biclustering pour identifier les modules biochimiques qui sont modulés par plusieurs produits chimiques in vivo et in vitro à travers les espèces.

Nous avons constaté que les produits chimiques induisent des motifs biochimiques conservés in vitro et in vivo chez l'homme et le rat. Ces modules transcriptionnels appartiennent aux voies de survie cellulaire, l'inflammation, le métabolisme des xénobiotiques, le stress oxydatif et l'apoptose. De plus, nos résultats confirment que la voie régulée par le récepteur de TGF-beta constitue un biomarqueur associé à l'exposition aux contaminants environnementaux dans les hépatocytes humains primaires.

Notre analyse intégrative des données de toxicogénomique fournit un aperçu complet sur les perturbations biochimiques affectées par un large panel de produits chimiques. En conclusion, nous avons démontré que la réponse toxicologique précoce survenant chez le rat, modèle privilégié en toxicologie, est reproduite dans des modèles cellulaires chez l'homme et le rat au niveau moléculaire. Ces résultats enrichissent notre compréhension et interprétation des

données de toxicogénomique dans les hépatocytes humains exposés à des toxiques environnementaux et par suite une extrapolation au test toxicologique de routine.

Contributions par auteur :
NEH et PG ont participé à la conception globale de l'étude, et sont responsables de la collecte des données et de la normalisation des échantillons de microarrays, l'analyse statistique, le code et l'implémentation du modèle en langage R, la rédaction du manuscrit, et l'interprétation des résultats. HJWLA et BHK conçu et supervisé l'étude et ont été responsables de l'interprétation des résultats. AB-C, ARB, NB et JA ont participé à la révision du manuscrit. Tous les auteurs ont lu et approuvé le manuscrit final.

Nehme El-Hachem[1, 2 $], Patrick Grossmann[3,6 $], Alexis Blanchet-Cohen[4], Alain R. Bateman[5], Nicolas Bouchard[2, 8], Jacques Archambault[9], Hugo J.W.L. Aerts[3, 6,7*] & Benjamin Haibe-Kains[10, 11*]
[1]Integrative systems biology, Institut de Recherches cliniques de Montréal, Montréal, Québec, Canada, Montréal, QC, Canada
[2]Department of Medicine, University of Montréal, Montréal, Canada
[3]Department of Biostatistics & Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA
[4]Bioinformatics, Institut de Recherches cliniques de Montréal, Montréal, Canada.
[5]Department of Human Genetics, McGill University, Montreal, Quebec, Canada.
[6]Departments of Radiation Oncology Dana-Farber Cancer Institute, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
[7]Department of Radiology, Dana-Farber Cancer Institute, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA,
[8]Molecular Biology of Neural Development, Institut de Recherches Cliniques de Montréal, Canada
[9]Laboratory of Molecular Virology, Institut de Recherches cliniques de Montréal, Montréal, QC, Canada
[10]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada;
[11]Medical Biophysics Department, University of Toronto, Toronto, Ontario, Canada

4.1 Abstract

Background : Genome-wide expression profiling is increasingly being used to identify transcriptional changes induced by drugs and environmental stressors. In this context the TG-GATEs project (Toxicogenomics Project-Genomics Assisted Toxicity Evaluation system)

generated transcriptional profiles from rat liver samples and human/rat cultured primary hepatocytes exposed to more than 100 different chemicals.

Objectives : To assess the capacity of the cell culture models to recapitulate pathways induced by chemicals *in vivo*, we leveraged the TG-GATEs dataset to compare the early transcriptional responses observed in the liver of rats treated with a large set of chemicals to those of cultured rat and human primary hepatocytes challenged with the same compounds *in vitro*.

Methods: We developed a new pathway-based computational pipeline that efficiently combines gene set enrichment analysis (GSEA) using Reactome pathways and biclustering to identify common modules of pathways that are modulated by several chemicals *in vivo* and *in vitro* across species.

Results: We found that chemicals induce conserved patterns of early transcriptional responses in *in vitro* and *in vivo* settings, and across human and rat. These responses involved pathways of cell survival, inflammation, xenobiotic metabolism, oxidative stress, and apoptosis. Moreover, our results support TGF-beta receptor signalling pathway as a candidate biomarker associated with exposure to environmental toxicants in primary human hepatocytes.

Conclusions : Our integrative analysis of toxicogenomics data provides a comprehensive overview of biochemical perturbations affected by a large panel of chemicals. Furthermore, we show that the early toxicological response occurring in animals is recapitulated in human and rat primary hepatocyte cultures at the molecular level, indicating that these models reproduce key pathways in response to chemical stress. These findings expand our understanding and interpretation of toxicogenomics data from human hepatocytes exposed to environmental toxicants.

4.2 INTRODUCTION

Humans are exposed to a variety of toxic chemicals and have access to a wide array of drugs each of which have the potential to cause short and long-term adverse effects including lethality. From an environmental health perspective, it is important to find a strong connection between toxic substances and human disease susceptibility, therefore elucidating molecular mechanisms of toxicity.

Although animal models are currently the gold standard in evaluating risk and predicting adverse human health effects, they require considerable time and resources, and

also raise ethical issues.[1-5]. For these reasons, several efforts have been made to minimize the use of animals in toxicology (http://www.alttox.org) and to develop robust *in vitro* models predictive of toxicity in human[6]. A European initiative, the REACH (Registration, Evaluation, Authorization and Restriction of chemicals) legislation, suggests the use of high-throughput "omics" technologies, such as genome-wide gene expression profiling, to find alternatives to animal testing. The REACH legislation states:

"The Commission, Member States, industry and other stakeholders should continue to contribute to the promotion of alternative test methods on an international and national level including computer supported methodologies, in vitro methodologies, as appropriate, those based on toxicogenomics, and other relevant methodologies" (REACH).

Multiple studies used gene expression profiles to characterize toxicogenomic response[7][8], as reviewed in[9]. To confront chemical induced cellular stress, the biological system executes a transcriptional control over several signaling pathways[10][11]. Because the liver plays a primordial role in detoxification and is a major site of frequent chemical-induced injuries, it was extensively studied in toxicogenomics. Recently, the Japanese government and the pharmaceutical industry joined forces to create and make publicly available the largest toxicogenomic database to date: the Toxicogenomics Project-Genomics Assisted Toxicity Evaluation system (TG-GATEs)[12][13]. The TG-GATEs consortium tested approximately 150 chemicals in different models, including primary human and rat hepatocytes as well as rat liver and kidney *in vivo* models[12][13]. The experimental design and gene expression profiles were made publicly available through the EBI ArrayExpress database http://www.ebi.ac.uk/arrayexpress/[14]. Different studies used this large toxicogenomic dataset to identify predictive biomarkers of hepatocarcinogenicity[15][16], phospholipidosis, and coagulopathy[17]. However, despite the availability of these valuable data, it remains unclear whether animal studies could be efficiently replaced by *in vitro* testing to identify key biological pathways induced by hepatotoxic chemicals, one of the main challenges of toxicogenomics.

In this study, we performed a large-scale comparative analysis of the TG-GATEs data from rat liver samples (referred to as RLV) and from cultured rat and human primary hepatocytes (referred to as PRH and PHH) in order to (*i*) identify conserved transcriptional responses induced by chemicals across species and between *in vitro* and *in vivo* systems, and

(*ii*) characterize the early response pathways linked to toxicity in both rat *in vivo* and rat/human *in vitro* experiments. Building upon the recent study of Iskar et al.[18] showing that drugs affected modules of co-expressed genes conserved across a small set of three human cancer cell lines and rat liver samples, we developed a new pathway-based approach that combined gene set enrichment analysis (GSEA) and biclustering to efficiently integrate large-scale toxicogenomic data across different species. Our analysis showed that chemicals affect a set of conserved pathways linked to chemical-induced toxicity across species and experimental platforms.

4.3 MATERIAL AND METHODS

The overall design of our analysis is represented in Figure 1. The three experimental settings that we investigated in TG-GATEs are rat liver *in vivo*, rat and human primary hepatocyte *in vitro* and are referred to as RLV, PRH and PHH, respectively.

4.3.1 Microarrays retrieval and preparation

Rat liver, primary rat and human hepatocytes microarray data files were downloaded from ArrayExpress. The three studies with the accessions E-MTAB-799, E-MTAB-798, and E-MTAB-797 contain toxicogenomic data for rat liver *in vivo* (RLV), cultured primary human hepatocytes (PRH), cultured primary human hepatocytes (PHH) and experiments, respectively, for more than 100 chemical compounds (Figure 1A). PHH and PRH were treated with each compound in duplicates, using three increasing doses (low, middle and high doses) for three different amounts of time (2, 8 and 24 hours; Figure 1A). Rat liver samples were obtained from animals treated with each compound in triplicates and sacrificed at 3, 6, 9 and 24 hours after dosing (Figure 1A). The highest dose refers to the maximally tolerated dose. Each compound is associated with a corresponding vehicle control for all experimental conditions.

All CEL files (Affymetrix data format that contains a signal for both perfect match and mismatch probes) were checked for duplicated names and inconsistencies. For 71 chemicals, it was noted that the data from human hepatocytes treated with a low dose of compound was missing; these 71 chemicals were nevertheless retained and analyzed with the other 48 chemicals. In total, the transcriptional effects of 119 chemicals on human hepatocytes were

gathered from 2,004 microarrays (Affymetrix GeneChip Human Genome U133 Plus 2.0 platform). Similarly, the effects of 129 chemicals on rat liver samples and rat hepatocytes were deduced from 6,192 and 3,096 microarrays, respectively (Affymetrix GeneChip Rat Genome 230_2.0) (Figure 1B). All datasets, including kidney samples in E-MTAB-799 and the repeated dose study (accession E-MTAB-800), are downloaded and curated on the fly through our fully automated pipeline. Documented code is available on GitHub (https://github.com/bhaibeka/TGGATES).

4.3.2 Gene expression data

Gene expression data were normalized with the robust multi-array average algorithm (RMA)[19] using the Bioconductor package BufferedMatrixMethods (version 1.30.0)[20]. Probes were mapped to Entrez Gene IDs using the Bioconductor annotation packages hgu133plus2.db (version 3.0.0) and rat2302.db (version 3.0.0) for human and rat, respectively. In case of multiple probes mapped to the same Entrez Gene ID, we used the Bioconductor package genefu (version 1.15.0) to select the most variant probeset for each gene. This procedure yielded 20,590 and 14,462 unique genes for human and rat, respectively.

4.3.3 Pathway collections

Every gene in the curated microarray experiments in TG-GATEs was assigned to pathways described in the Reactome database[21] using the Bioconductor package BiomaRt (version 2.22.0), for both rat and human genes present in the microarray platform. Pathway collection was performed on March 5, 2014. We subsequently selected the common pathways for rat and human, and retained only gene sets of sizes between 15 and 500 genes, which resulted in 419 common Reactome pathways for the GSEA analysis. For reproducibility, all curated pathways were stored in gmt files provided in https://github.com/bhaibeka/TGGATES.

4.3.4 Gene-chemical associations

Gene ranking was based on gene-chemical associations, which were identified by fitting linear models to estimate the effect of chemical dosage on gene expression levels controlled by treatment time and interaction between dosage and time. For each pair of gene *i* and chemical *j*, we used the following model

$$G_i = \beta_0 + \beta_1 D_j + \beta_2 T_j + \beta_3 D_j T_j \quad \text{(Equation 1)}$$

where G_i denotes the expression value of gene i, D_j is the dose of chemical j, T_j is the treatment time with chemical j, β_0 is the intercept, β_1, β_2, and β_3 are the regression coefficients for the chemical dosage, treatment time, and interaction term of dose and treatment, respectively. The strength of the gene-chemical association is given by β_1 and its significance (p) is computed using Student t test as provided by the lm() function in the R stats package (Team 2013).

4.3.5 Pathway-chemical associations

Pathways that are significantly perturbed by each chemical were identified using the java implementation of GSEA[22] (version 2.0.14) provided by the Broad Institute. For each chemical, we first ranked all genes with respect to the signed significance of their gene-chemical association, that is sign(β_1) * $-\log_{10}(p)$ as in (Equation 1). We then used each chemical-specific ranked list of genes to perform a pre-ranked GSEA to calculate normalized enrichment scores (NES) for all common pathways between human and rat. The higher the absolute value of NES, the more enriched is the corresponding pathway in genes whose expression is significantly perturbed by the chemical of interest. We repeated this process for each chemical and created an "enrichment matrix" with pathway enrichment scores in rows and chemicals in column for each dataset (Figure 4.1B).

4.3.6 Conserved transcriptional modules

One hundred and fifteen chemical compounds were common to all three experimental settings (Figure 1B). For each of these datasets, we applied a biclustering method, that is the iterative signature algorithm (ISA)[23] implemented in the R package isa2 (version 0.3.3)[24], on the enrichment matrix to simultaneously identify similar biochemical-induced transcriptional response patterns. The ISA algorithm runs with all combinations of threshold values on rows and columns, as described in details on the companion website (http://www.pmgenomics.ca/bhklab/pubs/tggates). Similarly to[18] we merged modules with similar set of pathways using function *isa.unique()* in the *isa2* package to filter redundant modules using a correlation limit of 0.5 to determine redundant biclusters. Lastly, modules sharing common sets of pathways and chemical across the different datasets -- namely RLV,

PRH, and PHH (inter-dataset similarity) -- were identified using a one-sided hypergeometric test ($p < 1E-3$); this technique is referred to as the reciprocal best-hit approach[18].

4.3.7 Reproducible research

To ensure full reproducibility, this work complies with the guidelines proposed by Robert Gentleman[25] in terms of availability of the code and reproducibility of results and figures. The procedure to properly set up the software environment and run our analysis pipeline is provided in Supplemental Material, Reproducibility of analysis. The analysis code is also publicly available on https://github.com/bhaibeka/TGGATES.

4.4 RESULTS

The approach we used to investigate the pathways altered by chemical perturbations leverages the transcriptional profiling data available in TG-GATEs for rat liver *in vivo* (RLV) and for rat and human primary hepatocytes cultured *in vitro* (PRH and PHH, respectively), as summarized in Figure 4.1A. We analyzed each of these three datasets separately and compared the results from the *in vitro* treated hepatocytes (PRH and PHH) to those from the liver of treated rats (RLV), as this animal model is considered the gold standard in toxicity studies. Pre-processing of these gene expression datasets yielded a set of 20,590 and 14,460 unique genes from the human and rat microarray platforms, respectively, that were kept for subsequent analysis. The association between gene expression levels and the 115 chemicals common across the three TG-GATEs experimental settings (PRH, PHH, and RLV) was then investigated at the pathway-level using the pre-ranked version of gene set enrichment analysis (GSEA)[22]. This was done with 419 pathways, which were in common between rat and human organisms as queried from Reactome database, in order to identify modulated pathways upon chemical perturbation. Matrices containing the enrichment scores of each pathway perturbed by each chemical were then analyzed using an unsupervised biclustering technique called Iterative Signature Algorithm (ISA)[23] to define functional modules (i.e., clusters of pathways) that are specifically associated with diverse chemical treatments. Each module is given a summary name, that is a Reactome parent term that best recapitulates the pathways enriched in this module.

4.4.1 Conservation of transcriptional modules across experimental settings

Rat liver in vivo (RLV) treated with a single dose: Twenty-four non-redundant modules were identified using the aforementioned ISA analysis ($p < 1E-3$). These modules were enriched for the following biological pathways: neuronal system, hemostasis, cell cycle checkpoints, DNA repair, mitosis, lysosomes disorders, innate immune system, NOTCH, TGF-βR/SMAD and PI3K/AKT signalling cascades, lipid metabolism, and mitochondrion dependant processes.

Primary Rat hepatocytes (PRH) vs. Rat Liver in Vivo (RLV): The ISA algorithm detected eighteen modules in PRH. Interestingly, seventeen modules overlapped with RLV using a reciprocal best-hit approach in which two modules are considered as conserved if their Reactome pathways significantly overlap (Iskar et al. 2013) (hypergeometric, $p <1E-3$). Only one module related to cholesterol biosynthesis did not overlap at the considered cutoff. Figure 4.2 shows in detail the number of non-redundant ISA modules in each dataset and the conservation across the experimental settings.

Primary human hepatocytes (PHH) vs. Rat Liver in Vivo (RLV): ISA analysis resulted in the identification of fifteen modules in PHH toxicogenomic data. Again, fourteen of them overlapped with RLV (hypergeometric, $p <1E-3$; Figure 4.2).

Overall, we identified thirteen modules to be conserved across the three experimental setting datasets (RLV, PHH, and PRH). (Figure 4.2). As a representative example, we show a conserved module in Figure 4.3. It is enriched for components of the innate immune system, with the overlapping pathways in the same order for both RLV, PHH and PRH. We extracted the union of the genes that were found to contribute to the enrichment score (referred to as *leading edge;* [Subramanian et al. 2005]) of at least one pathway for all chemicals in the module. From this, we obtained a list of common genes that are activated or repressed by chemical stress in RLV, PRH and PHH (Heatmaps for all ISA settings, lists of hypergeometric p-values, and lists of leading edge genes are provided in online Supplemental Material files https://www.pmgenomics.ca/bhklab/pubs/tggates).

4.4.2 Enrichment for hepatocarcinogens

The approach described above identified thirteen modules that are associated with the early response of hepatocytes to diverse chemicals and are conserved *in vivo*, *in vitro* and between rat and humans. To test if some modules were significantly associated with the hepatocyte

response to known hepatocarcinogens, we investigated twenty five previously validated rat hepatocarcinogens[16] present among the 115 chemicals investigated in our study. Specifically, these hepatocarcinogens were significantly enriched in the NOTCH and TGF-βR/SMAD signaling modules in PHH (hypergeometric $p < 0.05$), but not in PRH or RLV. The TGF-βR/SMAD signaling module (Figure 4.4A) in PHH was enriched for known environmental toxicants and carcinogens (e.g., ethionine, thioacetamide, coumarin, ethanol, acetamidofluorene, nitrosodiethylamine). None of these modules was enriched for hepatocarcinogens in RLV and this was only the case for the PI3K/AKT (Phosphoinositide 3-kinase) module in PRH ($p = 0.049$). The known rat hepatocarcinogens were also significantly associated with the neuronal system/G protein-coupled receptors (GPCRs) module in both RLV and PHH, but not PRH, probably reflecting the pleiotropic roles that GPCRs play in many cellular processes, including chemical carcinogenesis.

As a control experiment, we selected twelve non-carcinogenic compounds, and determined if they were significantly associated with any of the modules in RLV, PRH and PHH. As anticipated, no enrichment was observed, especially for those modules enriched for known hepatocarcinogens in PHH. As an additional control, we ascertained that the NOTCH and TGF-βR/SMAD modules were indeed enriched in cancer-related pathways; this was done by showing that the 20 pathways containing the word "Cancer" in the Reactome common dataset (over a total of 419 pathways) were in fact enriched in those modules (hypergeometric, $p < 1E-3$). This was not the case for any of the remaining modules without cancer terms. Collectively, the results presented above support that primary human hepatocytes can detect potential environmental chemical carcinogens (Figure 4.4A). By extension, we infer that the other modules are also enriched in pathways that are pertinent to chemical exposure.

4.4.3 Activation of the Peroxisome proliferator activated-receptor alpha (PPARalpha)

Since some PPARalpha activators are known to induce hepatocarcinogenesis in rodents' liver, we tested if PPARalpha activators (e.g. benziodarone, benzbromarone, fenofibrate, clofibrate, ibuprofen, WY14643, and gemfibrozil) were randomly distributed across modules in RLV, PHH and PRH. Interestingly, none of the modules in PHH or PRH were enriched for those

drugs, however we found that a module unique to RLV was significantly associated with the regulation of lipid metabolism by PPARalpha and enriched for those drugs ($p = 0.014$). Other PPARalpha potential inducers were found in this module including non-steroidal anti-inflammatory (NSAIDs) and anti-tuberculosis drugs (Figure 4.4B).

A recent study[10] showed that numerous compounds from TG-GATEs cause ”stereotypical” transcriptional responses in PHH. Such definition is given when a cytotoxic concentration of numerous compounds caused a consensus expression response regardless of the chemical class of compound. We assessed the significance of the overlap, for each module, between all leading edge genes, which we generated from the biclustering in PHH, and the deregulated genes by at least 20 compounds in their study. We demonstrated that “stereotypical” clusters of genes, involved in liver metabolic functions and cell proliferation, were enriched in two modules from PHH, mainly those associated with normal liver function and DNA synthesis modules. Furthermore, to ascertain that our observations from PHH are not simply experimental artefacts due to *in vitro* conditions, we selected liver cirrhosis as a case study and tested the enrichment for genes associated exclusively with liver cirrhosis in PHH[10]. Interestingly, the transforming growth factor beta-receptor signaling module in PHH (TGF-βR, module 6) was significantly enriched for genes linked to liver cirrhosis besides being induced by known hepatocarcinogens and environmental toxicants (Figure 4B).

Finally, we showed that the distribution of genes perturbed by rat hepatocarcinogens vs. non-hepatocarcinogens was alike.

4.5 DISCUSSION

We tested the extent to which transcriptional responses associated with liver toxicity can be recapitulated across human and rat and between *in vivo* and *in vitro* settings. To do so, we exploited the toxicogenomic information generated by the TG-GATEs project, from liver samples of rats treated with different chemicals and from rat/human hepatocytes exposed to the same compounds *in vitro*. To date, several studies have used TG-GATEs to build predictors of relevant toxicological endpoints. For example, Zhang et al. recently used this data to build a predictive gene signature for both hepatotoxicity and nephrotoxicity[26].

Interestingly, this study revealed the importance of early response genes in triggering toxicity-associated signalling networks, as highlighted by the high predictive power of the signature derived from a treatment period of less than 24 hours.

To our knowledge, our study is the first analysis of the TG-GATEs data comparing the functional changes - in the form of transcriptional responses - that are induced by a large panel of chemicals *in vivo* (rat liver)*, in vitro* (cultured hepatocytes) and between species (human vs. rat). A main feature of our approach is the fact that it relies on a pathway enrichment analysis, thus allowing comparison to be made between species without having to rely on the limited subset of orthologous genes. In this context, it is worth contrasting our findings to those of Iskar et al.[18], who identified, solely based on a orthologous genes, transcriptional modules that were conserved between rat liver and three human cancer cell lines from the Connectivity Map (CMap)[27]. Their findings showed that 15% of the chemical-induced modules were conserved across cell lines and species. However, this approach was limited to 8,962 genes in CMap, which corresponded to only 3,618 orthologous genes available for the rat liver experiments. To overcome this limitation, by focusing on common pathways between species, our approach enabled a full exploration of the TG-GATEs datasets and the identification of functional pathways altered by chemical treatments in both rat and human.

Our results indicate that the response of hepatocytes to chemical insults is analogous *in vitro*, *in vivo*, and across human and rat in that it involves a conserved set of cellular pathways. Specifically, we identified thirteen highly conserved modules representative of the early response of hepatocytes to chemical exposure. Two of those are enriched in key signalling pathways associated with cancer, namely the transforming growth factor beta receptor superfamily module (TGF-βR -mod17 RLV) the NOTCH signaling module (NOTCH-mod6 RLV). Given the role that the TGF-βR and NOTCH pathways play in response to early toxicity[26] and in maintaining normal liver functions[28], respectively, it was not surprising that these modules were enriched for known rat hepatocarcinogens including environmental toxicants. What could be more puzzling, according to our results, is the fact that these two pathways are significantly associated with hepatocarcinogens only in human and not in rat. This may reflect a key difference in how both species deal with these chemicals. That the response of rats and humans may differ for some chemicals is also supported by our finding that the PPARalpha agonists clofibrate, fenofibrate, gemfibrozil, benziodarone, and

benzbromarone up-regulate pathways associated with PPARalpha activation only in rat liver, thus providing a potential mechanism underlying the hepatocarcinogenicity of these drugs in rats but not humans[29].

Several lines of evidence suggest that the modules identified in this study are relevant to how hepatocytes respond to chemicals. For example, one of the modules we identified, the innate immune system (mod2 RLV), was enriched in pro-inflammatory Toll-like receptor signalling pathways, which have been shown by Huang et al. to be good predictors of drug induced liver injury[30]. Our results are also consistent with those reported in the comparative studies of Doktorova et al., who assessed the transcriptional profiles of toxicants between rat liver and a panel of *in vitro* models[31]. Those studies assigned deregulated genes from *in vivo/in vitro* comparison. Moreover, we found that pathways associated with G protein-coupled receptors (GPCRs) and the neuronal system were consistently affected by a variety of chemicals. Of particular relevance is the fact that some chemicals found in this conserved module (neuronal system-mod8 RLV) can cause the potentially lethal long QT syndrome (delayed repolarization of the heart) by perturbing heart conductance. For example, ciprofloxacin, haloperidol, thioridazine, quinidine, and amiodarone are well known to prolong the QT interval and cause *Torsades de Pointes,* a deadly form of arrhythmia[32]. This module was also enriched for known rat hepatocarcinogens in RLV and PHH but not PRH, a finding that may relate to the fact that ion channels, in addition to being involved in the long QT syndrome, can also play a role in carcinogenesis[33]. However, this observation might not be specific to a class of compounds since the Reactome pathways related to the neuronal system contain a large number of genes (> 500). Our findings also suggest that some chemicals modulate pathways associated with vitamin metabolism (metabolism of vitamins and cofactors-mod3 in RLV) in hepatocytes, in particular those associated with the inherited metabolic disorders ethylmalonic aciduria and homocystinuria. Surprisingly, the scientific literature contains only a few reports pertaining to the association between chemical-induced liver injury and vitamins. Amongst the studies that we found relevant to this work, one describes an association between high levels of circulating cobalamin (vitamin B12) and several serious liver diseases[34], while the other highlights the role of vitamin B12 metabolism in Methylmalonic aciduria, a disorder that can lead to severe liver injury and require in some cases a liver transplantation[35]. Given the strong association between vitamin

metabolism and early drug exposure revealed in our study, it may be of interest to explore further this understudied area of research.

Furthermore, we further confirmed the biological relevance of our biclusters against a recent study[10]. Indeed, we showed that our modules recapitulated stereotypic response to chemicals as well as compound-specific perturbations. Moreover, we found evidence that the TGF-β receptor signalling module in PHH could act as a potential biomarker of chemical injury that may lead to liver cirrhosis besides being enriched for known hepatocarcinogens.

It is worth mentioning that our new bioinformatics pipeline complements previous approaches, used to elucidate the mechanisms of chemical toxicity *in vitro* or *in vivo*, by enabling efficient and unbiased exploration of chemical-induced transcriptional changes in both *in vivo* and *in vitro* systems, and across different species. The modules that emerged from this analysis suggest that functional networks of xenobiotic detoxification and response to external stress are highly conserved in the hepatic system across human and rat. In contrast to pathway conservation, our results suggest that the chemicals associated with any given module, do not show a meaningful overlap between *in vitro* and *in vivo* systems or across species. Although somewhat counter-intuitive, this has been observed previously[26] and may reflect *bona fide* differences in chemical bioactivation through metabolism between systems, thus complicating the interpretation of *in vivo* versus *in vitro* data. Another factor to consider when assessing the value of our approach is the fact that it relied on an expert knowledge curated, peer-reviewed database of functional pathways. While it provided an alternative resolution for the orthologous gene limitation, we are nevertheless aware that annotations in pathway databases are incomplete and thus may limit this approach to some extent. Some of these limitations may be addressed in the future as we extend our approach to other systems (e.g., HepG2 hepatocellular carcinoma cell line), other toxicogenomic databases, such as DrugMatrix[36], and integrate more 'omics' data including RNA-seq.

4.6 Conclusion

The analysis of the TG-GATEs data presented here indicates that toxicogenomics-based cellular models recapitulate most of the pathways related to chemical-induced injury in rat liver. Furthermore, it may be possible to reduce unnecessary animal testing in early toxicological assessments and complement them with *in vitro* testing. Because environmental

toxicants can be associated with alterations in cellular pathways that contribute to general injury patterns and likely more severe phenotypes including carcinogenesis, we showed that the TGF-βR/SMAD module could serve as a putative biomarker to identify chemicals with carcinogenic potential for humans. Especially that potent carcinogenic compounds such as 2-acetamidofluorene, nitrosodiethylamine and ethanol were found in this module in PHH.

Our findings could be generalized to study a large set of environmental contaminants relevant to human health. Therefore, our method helps identify numerous pathways and genes that are responsible of toxicity controlled by chemical exposures.

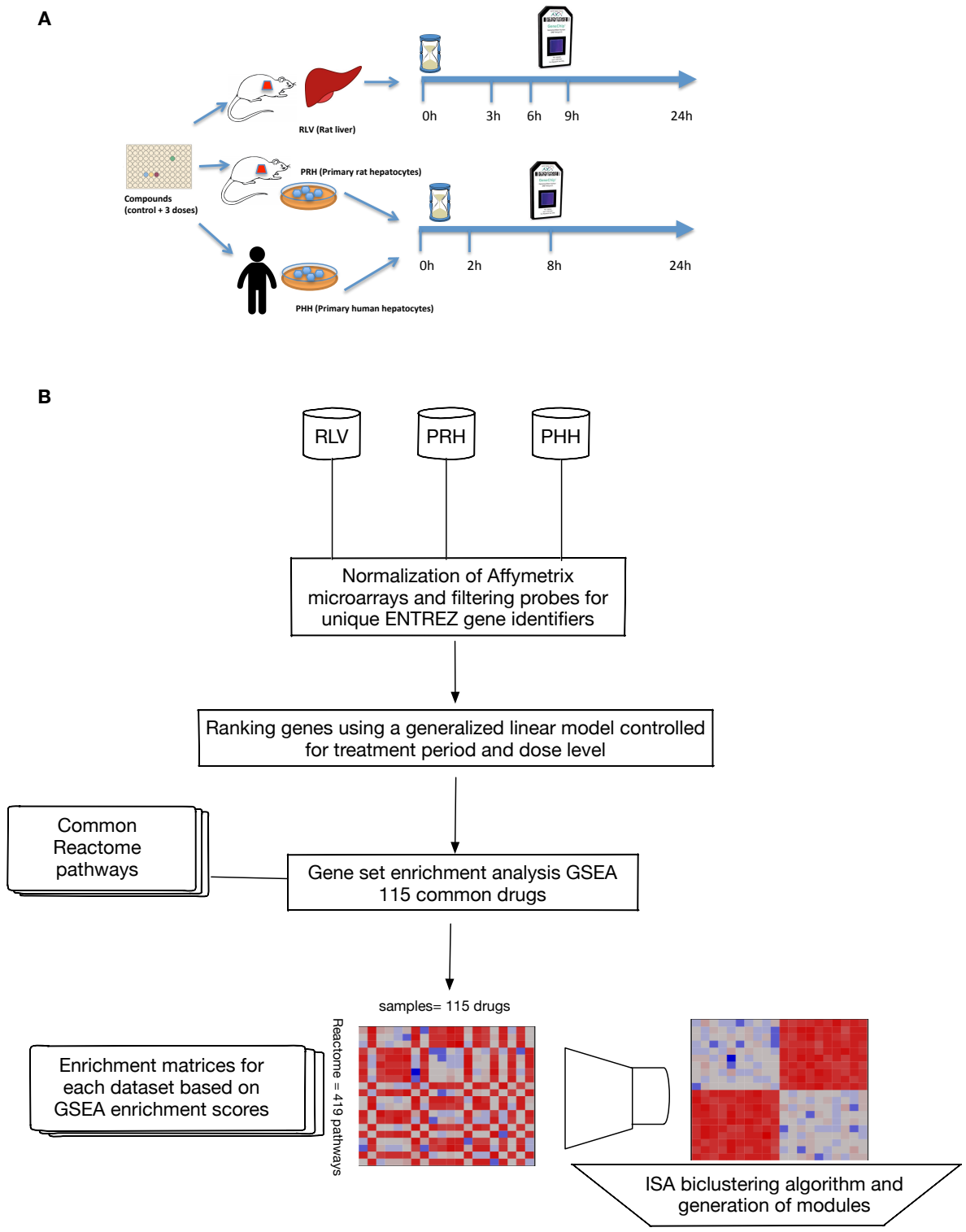

Figure 4.1 Analysis workflow for the TG-GATEs data set. (A) Overview of the TG-GATEs experimental design. TG-GATEs includes rat liver in vivo (RLV), rat hepatocyte in vitro (PRH), and human hepatocyte in vitro (PHH) experiments to test transcriptional responses for

> 100 chemical compounds. Samples have been treated with three doses of chemical alongside a control group, and gene expression was measured repeatedly within 24 hr (h) as shown. (B) Pathway-based analysis pipeline. A comparative analysis of the three TG-GATEs experiments was conducted by investigating chemical-induced pathways in RLV, PRH, and PHH. For each chemical, a linear regression model was fitted for every gene to assess the effects of the chemical on gene expression, taking into account the treatment period and the dose. Based on these association models, genes were ranked to perform a gene set enrichment analysis (GSEA) on common Reactome pathways. From the enrichment results, transcriptional modules conserved across experimental settings (RLV, PRH, and PHH) were identified by biclustering.

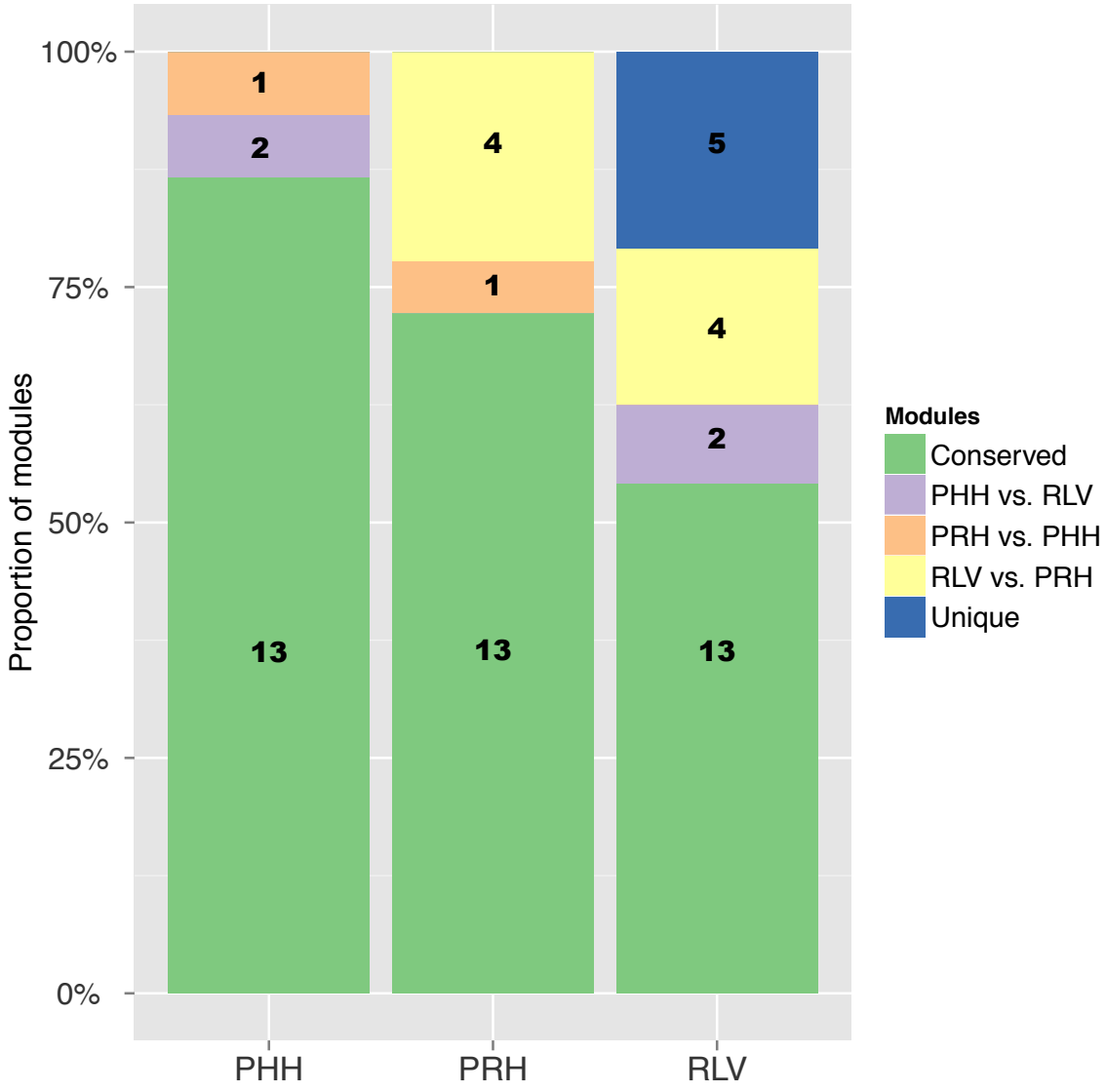

Figure 4.2 The number of non-redundant transcriptional modules and proportions identified for each and across all experimental settings in TG-GATEs. Each bar corresponds to an experimental setting in TG-GATEs (RLV, PRH, PHH) and contains the number of modules found to be unique for the experiment (blue) or shows a corresponding module in another experiment (see color legend: green for conserved, purple for conserved in RLV vs. PHH, orange for conserved in PHH vs. PRH, yellow for conserved in RLV vs. PRH). Although only a few modules were detected in only one or two settings, most modules showed significantly high overlap in terms of pathway enrichment across all settings (hypergeometric $p < 0.001$).

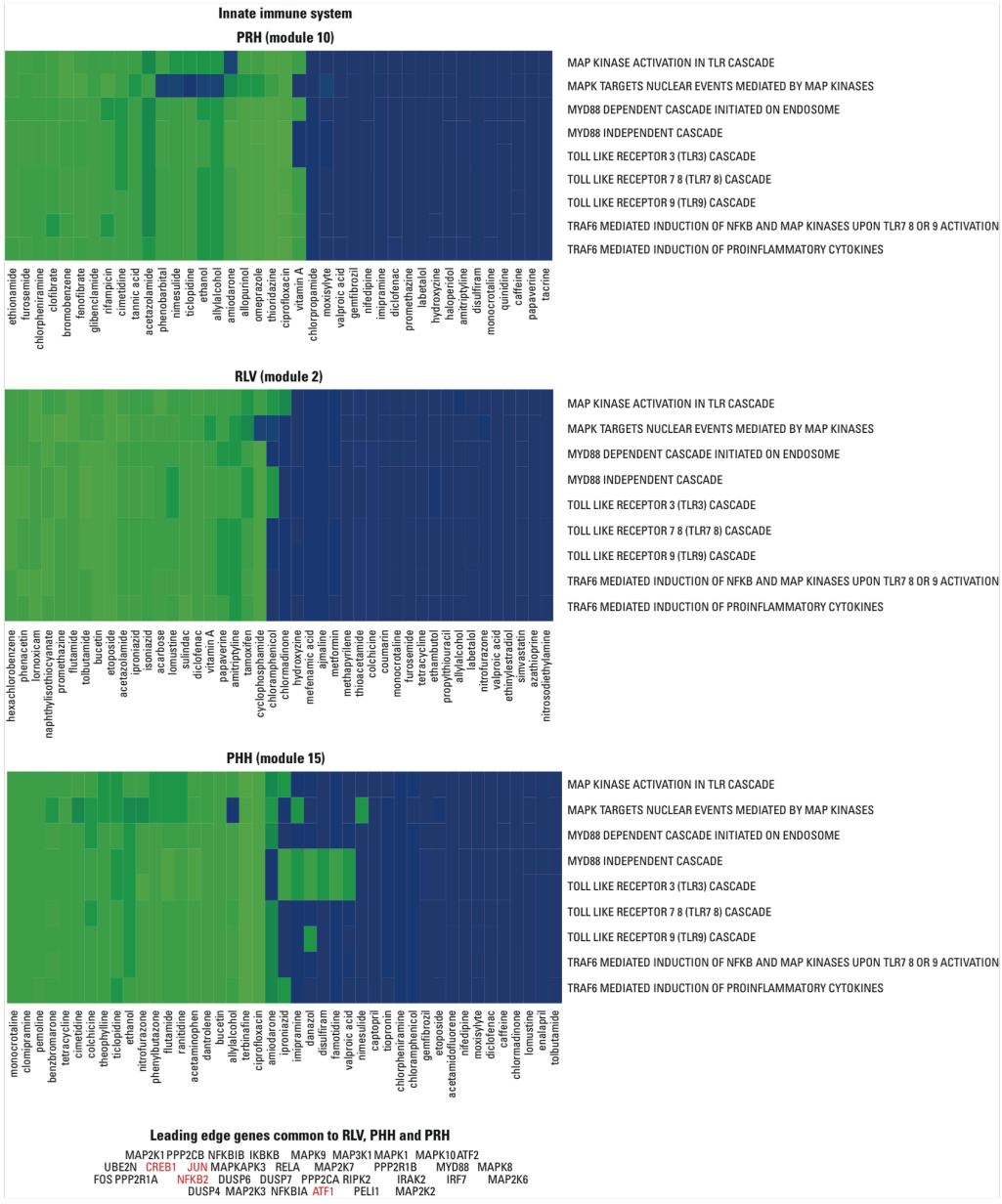

Figure 4.3 Conservation of modules across in vitro and in vivo settings based on Reactome pathways. This example summarizes a conserved module between RLV, PRH, and PHH, shown as heatmaps and keeping overlapping pathways colored with respect to their enrichment scores: up-regulated pathways are shown in blue, and down-regulated pathways are shown in green. The three heatmaps correspond to a conserved module associated with the innate immune system (mod2 in RLV, mod15 in PHH, and mod10 in PRH). The leading edge genes from common pathways that are activated or repressed by chemicals are shown under the heatmap with known oncogenes colored in red.

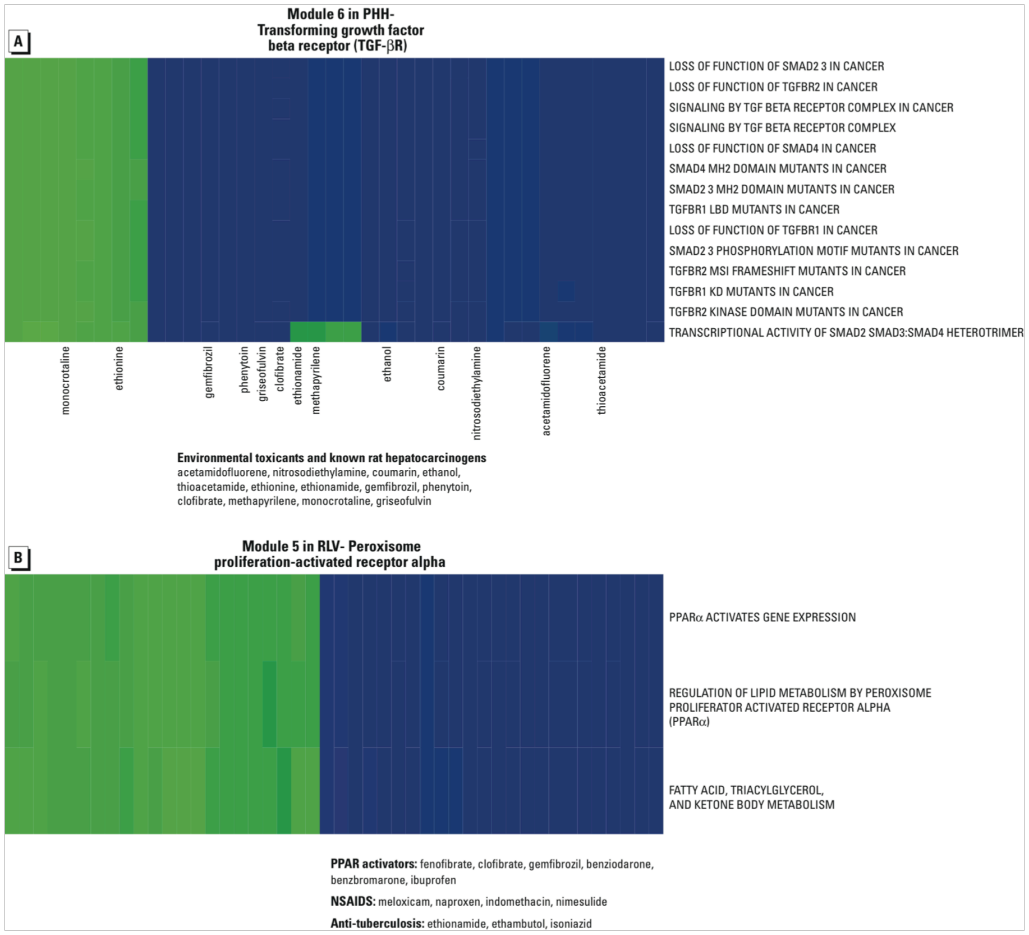

Figure 4.4 Characterization of putative biomarkers within chemical-induced modules. (A) Heatmap representing a module in PHH (mod6), associated with transforming growth factor beta receptor signalling, that can be considered as a candidate biomarker in humans for environmental exposure to known toxicants. Diverse rat hepatocarcinogens were enriched in this module. (B) Heatmap representing a module in RLV (mod5) that was relevant to toxicity mode of action and is enriched for a class of lipid-lowering drugs known as fibrates. These

drugs are rat hepatocarcinogens and activate the peroxisome proliferation-activated receptor alpha (PPARα).Green, down -regulation; blue, up-regulation. Drugs that activate PPAR pathways include nonsteroidal anti-inflammatory and antituberculosis drugs.

References

1. Bissell DM, Gores GJ, Laskin DL, Hoofnagle JH. Drug-induced liver injury: Mechanisms and test systems. Hepatology. Wiley Online Library; 2001; 33 : 1009–1013.
2. Greaves P, Williams A, Eve M. First dose of potential new medicines to humans: how animals help. Nat Rev Drug Discov. 2004; 3 : 226–236.
3. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? Nat Rev Drug Discov. 2004; 3 : 711–715.
4. Metushi IG, Uetrecht J. Lack of liver injury in Wistar rats treated with the combination of isoniazid and rifampicin. Mol Cell Biochem. 2014; 387 : 9–17.
5. Suter L, Schroeder S, Meyer K, Gautier J-C, Amberg A, Wendt M, et al. EU framework 6 project : predictive toxicology (PredTox)--overview and outcome. Toxicol Appl Pharmacol. 2011; 252 : 73–84.
6. Abbott A. Animal testing : more than a cosmetic change. Nature. 2005; 438 : 144–146.
7. Afshari CA, Nuwaysir EF, Barrett JC. Application of complementary DNA microarray technology to carcinogen identification, toxicology, and drug safety evaluation. Cancer Res. 1999; 59 : 4759–4760.
8. Ellinger-Ziegelbauer H, Gmuender H, Bandenburg A, Ahr HJ. Prediction of a carcinogenic potential of rat hepatocarcinogens using toxicogenomics analysis of short-term in vivo studies. Mutat Res. 2008; 637 : 23–39.
9. Afshari CA, Hamadeh HK, Bushel PR. The evolution of bioinformatics in toxicology: advancing toxicogenomics. Toxicol Sci. 2011; 120 Suppl 1 : S225–37.
10. Grinberg M, Stöber RM, Edlund K, Rempel E, Godoy P, Reif R, et al. Toxicogenomics directory of chemically exposed human hepatocytes. Arch Toxicol. 2014; 88 : 2261–2287.

11. Kier LD, Neft R, Tang L, Suizu R, Cook T, Onsurez K, et al. Applications of microarrays with toxicologically relevant genes (tox genes) for the evaluation of chemical toxicants in Sprague Dawley rats in vivo and human hepatocytes in vitro. Mutat Res. 2004; 549 : 101–113.

12. Uehara T, Ono A, Maruyama T, Kato I, Yamada H, Ohno Y, et al. The Japanese toxicogenomics project: application of toxicogenomics. Mol Nutr Food Res. 2010; 54 : 218–227.

13. Uehara T, Minowa Y, Morikawa Y, Kondo C, Maruyama T, Kato I, et al. Prediction model of potential hepatocarcinogenicity of rat hepatocarcinogens using a large-scale toxicogenomics database. Toxicol Appl Pharmacol. 2011; 255 : 297–306.

14. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, et al. ArrayExpress--a public repository for microarray gene expression data at the EBI. Nucleic Acids Res. 2003; 31 : 68–71.

15. Caiment F, Tsamou M, Jennen D, Kleinjans J. Assessing compound carcinogenicity in vitro using connectivity mapping. Carcinogenesis. 2014; 35 : 201–207.

16. Yamada F, Sumida K, Uehara T, Morikawa Y, Yamada H, Urushidani T, et al. Toxicogenomics discrimination of potential hepatocarcinogenicity of non-genotoxic compounds in rat liver. J Appl Toxicol. 2013; 33 : 1284–1293.

17. Hirode M, Ono A, Miyagishima T, Nagao T, Ohno Y, Urushidani T. Gene expression profiling in rat liver treated with compounds inducing phospholipidosis. Toxicol Appl Pharmacol. 2008; 229 : 290–299.

18. Iskar M, Zeller G, Blattmann P, Campillos M, Kuhn M, Kaminska KH, et al. Characterization of drug-induced transcriptional modules: towards drug repositioning and functional understanding. Mol Syst Biol. 2013; 9 : 662.

19. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003; 4 : 249–264.

20. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004; 20 : 307–315.

21. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, et al. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 2005; 33 : D428–32.

22. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005; 102 : 15 545–15 550.

23. Bergmann S, Ihmels J, Barkai N. Iterative signature algorithm for the analysis of large-scale gene expression data. Phys Rev E Stat Nonlin Soft Matter Phys. 2003; 67 : 031902.

24. Csárdi G, Kutalik Z, Bergmann S. Modular analysis of gene expression data with R. Bioinformatics. 2010; 26 : 1376–1377.

25. Gentleman R. Reproducible research: a bioinformatics case study. Stat Appl Genet Mol Biol. 2005; 4 : Article2.

26. Zhang JD, Berntenis N, Roth A, Ebeling M. Data mining reveals a network of early-response genes as a consensus signature of drug-induced in vitro and in vivo toxicity. Pharmacogenomics J. 2014; 14 : 208–216.

27. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006; 313 : 1929–1935.

28. Morell CM, Strazzabosco M. Notch signaling and new therapeutic options in liver disease. J Hepatol. 2014; 60 : 885–890.

29. Lai DY. Rodent carcinogenicity of peroxisome proliferators and issues on human relevance. J Environ Sci Health C Environ Carcinog Ecotoxicol Rev. 2004; 22 : 37–55.

30. Huang J, Shi W, Zhang J, Chou JW, Paules RS, Gerrish K, et al. Genomic indicators in the blood predict drug-induced liver injury. Pharmacogenomics J. 2010; 10 : 267–277.

31. Doktorova TY, Ellinger-Ziegelbauer H, Vinken M, Vanhaecke T, van Delft J, Kleinjans J, et al. Comparison of genotoxicant-modified transcriptomic responses in conventional and epigenetically stabilized primary rat hepatocytes with in vivo rat liver data. Arch Toxicol. 2012; 86 : 1703–1715.

32. Fazio G, Vernuccio F, Grutta G, Re GL. Drugs to be avoided in patients with long QT syndrome: Focus on the anaesthesiological management. World J Cardiol. 2013; 5 : 87–93.

33. Babcock JJ, Li M. hERG channel function: beyond long QT. Acta Pharmacol Sin. 2013; 34 : 329–335.

34. Ermens AAM, Vlasveld LT, Lindemans J. Significance of elevated cobalamin (vitamin B12) levels in blood. Clin Biochem. 2003; 36 : 585–590.

35. Hansen K, Horslen S. Metabolic liver disease in children. Liver Transpl. 2008; 14 : 713–733.

36. Natsoulis G, Pearson CI, Gollub J, P Eynon B, Ferng J, Nair R, et al. The liver pharmacological and xenobiotic gene response repertoire. Mol Syst Biol. 2008; 4 : 175.

CHAPITRE 5 : DISCUSSION ET CONCLUSIONS

5.1 Préambule

Une des motivations premières de cette thèse était de mieux comprendre la relation entre les déterminants génomiques/trancriptomiques/pharmacologiques et le mode d'action des médicaments (e.g. biomarqueur prédictif, cible moléculaire, toxicité), et ceci en vue d'évaluer les avantages et limitations du modèle de lignée cellulaire dans le processus du développement des médicaments. Cette thèse a pour but de développer des approches analytiques afin d'exploiter les grandes bases de données pharmacogénomiques et toxicogénomiques. Nos diverses contributions méthodologiques peuvent se résumer comme suit :

1. Nous avons montré qu'il existe une discordance entre deux grandes études pharmaco génomique, CCLE et GDSC, pour 15 médicaments et 470 lignées cellulaires en commun. Cette discordance était surtout due aux données pharmacologiques (essais cellulaires pour déterminer la concentration qui élimine 50 % de la population cellulaire). Ceci indique que la standardisation des essais pharmacologiques est indispensable pour identifier des biomarqueurs de réponse robustes et reproductibles. Nous avons montré que dans la plupart des cas, une analyse basée sur des ensembles de gènes représentant un processus biologique, à la place des gènes individuels, est plus robuste pour développer des biomarqueurs prédictifs de la réponse aux médicaments. Nous avons aussi montré que la mesure AUC (aire sous la courbe pharmacologique) est meilleure que IC_{50}. Notre étude originale a déclenché une nouvelle tendance dans le milieu de recherche sur la pharmacogénomique avec des études récentes basées sur notre travail (Cancer Cell Line Encyclopedia Consortium & Genomics of Drug Sensitivity in Cancer Consortium, 2015; Haverty et al., 2016; Pozdeyev et al., 2016).
2. Nous avons développé une méthode intégrative qui exploite plusieurs types de données pour le médicament : la structure chimique, les mesures de sensibilité cellulaire aux médicaments (dans les lignées cancéreuses NCI60 ou CTRPv2), l'expression de ~1000 gènes perturbée par le médicament à partir de la base de données LINCS L1000. Ainsi,

nous avons mis en évidence que l'intégration de ces trois niveaux de données était significativement plus performante à prédire le mode d'action des médicaments que si l'on considère chaque niveau individuel, ce qui est le cas dans la plupart des études précédentes.

3. Nous avons montré qu'il existe des signatures de toxicités communes interespèces (homme/rat) et que les modèles cellulaires hépatiques pourraient récapituler en partie les mécanismes de toxicité induits par les médicaments *in vivo*. En plus, notre étude toxicogénomique intégrative a permis de caractériser des voies de signalisation induites par des carcinogènes, exclusivement dans les cellules hépatiques humaines. Notre étude a permis de caractériser un répertoire de mécanismes moléculaires qui sont déclenchés par l'exposition a différents types de produits chimiques.

5.1.1 Biomarqueurs de réponse aux médicaments et concordance entre les études pharmacogénomiques

Notre analyse de ces études pharmacogénomiques pointe vers un problème fondamental dans l'évaluation de la réponse pharmacologique du médicament. Bien que l'analyse de l'expression des gènes a longtemps été considérée comme une source de variabilité, d'importants travaux ont conduit à des approches bioinformatiques robustes et standardisées pour la collecte, la normalisation et l'analyse des données, ainsi que le développement de plates-formes robustes pour mesurer les niveaux d'expression. Cette standardisation des technologies a conduit à des ensembles de données d'expression plus reproductibles et de qualité plus élevée, et cela est évident dans CCLE et GDSC où nous avons trouvé une excellente corrélation entre les profils d'expression dans les mêmes lignées cellulaires profilées dans les deux études.

Par contre, la faible corrélation entre les profils de réponse pharmacologique aux anticancéreux est troublante et peut représenter un haut niveau de variation technique et biologique, ainsi qu'un manque de standardisation dans les essais expérimentaux et les méthodes analytiques d'extrapolation des mesures de viabilité cellulaires. Cependant, nous avons montré que l'utilisation de différents protocoles expérimentaux ne peut être la seule source d'inconsistance. En effet nous avons observé une corrélation faible à modérée

(corrélation de perason < 0,6) entre les mesures de réponse (IC50) à la camptothécine et AZD6482, générées dans deux sites différents, en utilisant la même collection de lignées cellulaires et protocoles expérimentaux identiques (Safikhani et al., 2016). Il est improbable que les lignées cellulaires puissent être la cause principale des différences phénotypiques observées, car les profils d'expression génique sont bien corrélés entre les études.

Bien que notre analyse fût limitée à des lignées cellulaires et médicaments en commun entre les études, il soit raisonnable de supposer que la réponse mesurée pour les autres médicaments et lignées cellulaires testées est également discutable. En fin de compte, la faible corrélation dans ces études publiées constitue un obstacle à l'utilisation des ressources pharmacogénomique (surtout liées aux études du cancer) pour identifier ou valider des déterminants génomiques/transcriptomiques prédictifs de la réponse aux médicaments. Parce qu'il n'y a pas de concordance claire, les modèles prédictifs développés à partir des données d'une étude sont presque garantis à l'échec lorsqu'ils sont validés sur des données provenant d'une autre étude, comme nous l'avons montré dans notre étude publiée dans JAMIA en 2013 (Papillon-Cavanagh et al., 2013a) en accordance avec plusieurs études indépendantes qui ont suivies (Cortés-Ciriano et al., 2015; Z. Dong et al., 2015; Jang et al., 2014), Cela suggère qu'il faudra être prudent lors de l'interprétation des résultats issus de ces deux études.

Il est clair que l'investissement dans ces projets garantit un travail supplémentaire pour résoudre les divergences dans la réponse aux médicaments de telle sorte que la richesse des données générées peut être exploitée pour avancer la recherche sur le cancer. Nos résultats confirment la nécessité d'une standardisation des essais, ou le développement de nouvelles techniques uniformes et robustes. Sans cela, il sera imprudent d'identifier de manière fiable les prédicateurs génomiques de réponse à un médicament ou de déterminer efficacement le mécanisme d'action.

Nos résultats publiés ont provoqué de nombreuses réactions dans la communauté biomédicale et nous avons reçu un retour constructif mettant en avant des erreurs (mineures) dans notre étude et amélioration potentielles. Ce retour nous a permis de revisiter notre analyse initiale et de confirmer les résultats et conclusions publiés en 2013 (Safikhani et al., 2015). Nous abordons ci-dessous quelques-uns de ces problèmes et améliorations :

- **L'absence de quelques données de sensitivité :** Nous avons considéré dans notre étude que les lignées cellulaires qui ont été profilées pour l'expression génique. Or, il

nous a été pointé que quelques lignées cellulaires avaient des profils pharmacologiques, mais pas de profil transcriptomique, et étaient donc manquantes dans notre étude initiale. Nous avons corrigé notre erreur et inclut toutes les lignées cellulaires en commun entre les études (une trentaine supplémentaire).

- **La présence de plusieurs classes de médicaments :** Dans notre étude publiée, nous aurions dû stratifier notre analyse par les différentes classes de médicaments : (1) les médicaments avec aucune activité observée dans aucune des lignées cellulaires (2) les médicaments présentant une réponse observée pour seulement un ensemble de lignées cellulaires, et (3) les médicaments produisant une réponse dans un grand nombre de lignées cellulaires. Pour chaque classe, nous avons réévalué la corrélation de la réponse entre les deux études en utilisant une variété de paramètres statistiques. Même si aucune méthode n'a pu trouver de consistance pour la première classe de médicaments (sorafenib, erlotinib et PHA-665752), il était possible d'améliorer la corrélation en utilisant le bon choix de la méthode statistique pour la deuxième classe (nilotinib, crizotinib) et enfin cette corrélation est inchangée pour des agents de la troisième classe (par rapport a notre étude de 2013).
- **Le cas des thérapies ciblées et corrélation de spearman :** la corrélation de spearman, utilisée dans notre étude initiale, n'est pas sensible aux cas spéciaux (« outliers ») et donc n'est pas une mesure appropriée pour les thérapies ciblées. En effet, dans le cas des thérapies ciblées, la grande majorité des lignées cellulaires sont insensibles alors que quelques une arborant une aberration génomique donnée sont extrêmement sensibles. C'est le cas par exemple de nilotinib ou seulement 3 lignées cellulaires avec la fusion BCR-ABL1 ont montré une grande sensibilité, et ce, à la fois dans GDSC et CCLE. Cependant, la corrélation de spearman, ne considérant que le « rang » des lignées cellulaires, est proche de 0 malgré la concordance entre les études. Nous avons donc utilisé d'autres métriques de concordance et avons dès lors classifié les données de sensitivité au nilotinib comme étant concordantes. Malheureusement, les autres thérapies ciblées n'ont pas présenté une telle amélioration de concordance avec l'utilisation des autres métriques.

Nos résultats ont suscité une réponse des auteurs de CCLE et GDSC, qui, au lieu de répondre directement à notre étude, ont choisi de publier un manuscrit indépendant dans lequel

ils rapportent une concordance « raisonnable » entre leurs études respectives (Cancer Cell Line Encyclopedia Consortium & Genomics of Drug Sensitivity in Cancer Consortium, 2015). Bien que le titre de leur analyse comparative suggère une contradiction totale avec nos résultats, les auteurs ont reconnu que la cohérence des données pharmacologiques est loin d'être parfaite en raison des différences méthodologiques entre les protocoles utilisés par CCLE et GDSC, indiquant en outre que la standardisation de ces protocoles va certainement améliorer les mesures de corrélation. Nous avons évidemment investigué leur analyse dans les moindres détails et avons observé des différences fondamentales entre notre étude comparative initiale et l'évaluation récente publiée par les auteurs de GDSC et CCLE (Safikhani et al., 2016). Nous avons remis en question les choix analytiques des équipes de GDSC et CCLE qui, en plus d'être non justifiés, ont très probablement contribué à une concordance artificiellement élevée des biomarqueurs entre les deux études.

Malgré les avancées technologiques en pharmacogenomique du cancer, la validation des biomarqueurs de réponse aux anticancéreux, en se basant sur des modèles de lignées cellulaires, est pour l'instant un grand défi pour la médecine de précision. D'où la nécessité d'optimiser et standardiser les protocoles pharmacologiques pour de meilleures corrélations entre les études indépendantes. Toutefois, ces études sont susceptibles de rester une source préclinique cruciale pour la génération d'hypothèses sur le mécanisme des médicaments et leurs cibles moléculaires potentielles. Notre étude fut la première à comparer de larges études de pharmacogénomiques et a motivé de nombreuses études indépendantes cherchant à isoler les facteurs expérimentaux les plus cruciaux et à améliorer les analyses de ces données complexes (Cancer Cell Line Encyclopedia Consortium & Genomics of Drug Sensitivity in Cancer Consortium, 2015; Haverty et al., 2016; Pozdeyev et al., 2016).

5.1.2 Approche intégrative et prédiction du mécanisme d'action des médicaments

Plusieurs approches ont été proposées pour caractériser le MoA des médicaments. Ici, nous discuterons notre nouvelle méthode qui intègre les données pharmacogénomiques (voir chapitre 3). L'identification du/des MoA pour les nouveaux médicaments est un défi majeur, car ceci vise à caractériser les cibles primaires responsables de l'effet pharmacologique et les

hors-cibles « off-targets » associées à des effets physiologiques inattendus. L'une des principales limitations des approches actuelles est leur dépendance aux annotations pharmacologiques, biochimiques et thérapeutiques prédéfinies qui se rapportent à des médicaments caractérisés et approuvés pour des applications cliniques, et qui ne peut être applicable à une nouveau médicament (Ma'ayan et al., 2014; Napolitano et al., 2013).

Notre analyse aborde ces questions en menant, à notre connaissance, la première étude intégrative à grande échelle qui fusionne la similarité structurale et les caractéristiques pharmacogénomiques de base telles que les mesures de sensibilité aux médicaments dans les lignées cellulaires cancéreuses (NCI60 et CTRPv2) et les signatures transcriptomiques de ~1000 gènes perturbés par les médicaments dans une collection de lignées cellulaires (LINCS L1000). Nous avons adapté la méthode intégrative (SNF : Similarity Network Fusion, (B. Wang et al., 2014)) pour construire un réseau de similarité pour les médicaments (DNF : Drug Network fusion), basé sur la fusion de la structure, la sensibilité, et les données de perturbation. La fusion de ces trois types de données nous permet de générer une classification/taxonomie précise. Nos résultats ont indiqué que DNF surpasse chaque niveau pris individuellement (structure, ou sensibilité ou perturbation) pour la prédiction de la cible moléculaire et classification thérapeutique (ATC). Par conséquent, notre approche intégrative a réussi à combiner plusieurs types de données en un seul réseau global qui représente les caractéristiques fondamentales et expérimentales des médicaments. Ces types de données, par rapport à d'autres annotations y compris ATC, sont beaucoup plus faciles à générer pour des nouveaux médicaments non caractérisés. Notre taxonomie DNF a révélé l'importance de l'information issue des mesures de sensibilité cellulaire aux médicaments (NCI60 ou CTRPv2) en vue d'améliorer la performance de la prédiction des associations médicament cible. Ces résultats confirment et soulignent l'importance de l'ensemble de données CTRPv2 (860 lignées cellulaires) pour générer des prédicateurs de la réponse aux anticancéreux en plus de caractériser le mécanisme d'action (Seashore-Ludlow et al., 2015b).

Comme preuve de principe, la taxonomie DNF a classifié correctement toutes les classes de médicaments connues. Ces cas sont des contrôles positifs qui constituent la validation expérimentale de notre méthode. Nous avons classé correctement tous les inhibiteurs de BRAF (mutation V600E), qui comprennent les médicaments déjà testés dans le mélanome métastatique (dabrafenib, GDC0879, PLX4720), et les inhibiteurs de MEK

(mitogen-activated protein kinase), à savoir trametinib et selumetinib. BRAF régule la voie de signalisation MAPK/ERK hautement conservée, et le statut mutationnel de BRAF a été proposé comme un biomarqueur de sensibilité envers la médicament selumetinib et d'autres inhibiteurs de MEK (Solit et al., 2006). Nous avons également classé correctement les médicaments sans cibles/mécanismes annotés dans CTRPv2 tels que ifosfamide, cyclophosphamide et procarbazine qui sont connus comme agents alkylants (code ATC : L01A). Cela était également vrai aussi pour le docétaxel et le paclitaxel, deux taxanes (code ATC : L01CD).

Notre taxonomie a également consolidé les résultats antérieurs pour les médicaments qui peuvent servir de dérégulateurs de la polymérisation de la tubuline, et qui ne l'ont pas été précédemment classées comme telles. Nous avons identifié une communauté de trois médicaments : LY2183240 et YK-4-279 ont été récemment identifiés pour réduire les niveaux d'alpha-tubuline (Seashore-Ludlow et al., 2015b). Tivantinib, un inhibiteur de tyrosine kinase c-MET, a également bloqué la polymérisation des microtubules (Katayama et al., 2013). Aussi, cette communauté est étroitement liée aux perturbateurs connus de microtubules.

Nos résultats sont aussi concordants avec l'étude de Rees et al. (Rees et al., 2016) concernant les inhibiteurs de BCL-2 : ABT-737 et navitoclax, où les auteurs ont rapporté qu'une forte expression de BCL-2 confère une sensibilité à ces deux médicaments. Ce ne fut pas le cas pour un autre inhibiteur de BCL-2, obatoclax. Ils ont proposé qu'une modification métabolique du médicament obatoclax réduise son interaction avec la protéine Bcl-2. Nous avons montré en effet que obatoclax n'est pas dans le même cluster que les deux autres inhibiteurs de BCL-2 (ABT-737 et navitoclax). Un tel exemple montre comment les profils de sensibilité de ces deux inhibiteurs sont en grande partie cohérente contrairement à obatoclax, qui a montré précédemment des effets hors cible par rapport à ABT-737 (Vogler et al., 2009). Ceci fournit une bonne preuve pour examiner les profils de sensibilité lors de l'élaboration de nouveaux inhibiteurs spécifiques de BCL-2.

Nos résultats suggèrent l'existence de « super communautés », c'est-à-dire est un regroupement de plusieurs communautés qui contribuent à un MoA similaire, concordant avec des études précédentes menées avec CMAP, Un exemple est fourni par les communautés suivantes qui sont étroitement liées : les inhibiteurs bien caractérisés des Kinases cycline-dépendantes (CDK) qui sont connues pour être les principaux régulateurs du cycle cellulaire.

Ces composés sont positionnés à proximité des inhibiteurs de la topoisomérase I et II (SN-38, topotécan, étoposide, téniposide), les perturbateurs des microtubules (paclitaxel, docétaxel, vincristine, parbendazole) et les inhibiteurs de la kinase polo-like (GSK461364, GW843682X). Iorio et al., ont rapporté que la similarité entre les inhibiteurs de CDK et les autres agents endommageant l'ADN est médiée par une induction de p21, ce qui explique l'interconnexion de ces régulateurs de la progression du cycle cellulaire (Iorio et al., 2010).

Nos résultats récapitulent aussi les résultats d'essais cliniques. Par exemple, ibrutinib, qui est un inhibiteur de tyrosine kinase de Bruton (BTK) approuvé pour le traitement du lymphome des cellules du manteau et de la leucémie lymphoïde chronique, cluster avec les inhibiteurs connus de l'EGFR (erlotinib: géfitinib, afatinib). L'effet de l'ibrutinib dans le cancer du poumon a été rapporté dans un récent essai clinique (ClinicalTrials.gov Identifier : NCT02321540). Ce fut aussi le cas pour le MGCD265, un inhibiteur de MET, qui se groupe avec la plupart des inhibiteurs de VEGFR (pazopanib, cediranib). Dans cette communauté, pazopanib est le seul médicament approuvé par la FDA pour le traitement du carcinome rénal. Il existe une preuve récente que le MGCD265 a une application dans de nombreux types de cancers, y compris le carcinome rénal (ClinicalTrials.gov Identifier : NCT00697632).

Essentiellement, notre étude suggère que les données de sensibilité aux médicaments sont un atout important pour prédire le mécanisme du médicament. De plus, pour tester la robustesse de l'algorithme de fusion, nous avons également appliqué notre méthode sur l'ensemble de données NCI60, qui comprend un nombre beaucoup plus petit de lignées cellulaires (60 pour NCI60 contre 860 dans CTRPv2). Mais NCI60 compense son nombre limité de lignées par le grand nombre de médicaments testées (> 40.000). Certaines des communautés identifiées en utilisant NCI60, comme les inhibiteurs de BRAF/MEK, avaient également été identifiées dans notre analyse originale en utilisant des profils de sensibilité CTRPv2. Cela démontre un haut degré de spécificité des associations médicament cible dans des lignées cellulaires et entre les différentes plateformes expérimentales pour certaines classes de médicaments, ce qui est crucial pour l'identification des biomarqueurs et la recherche translationnelle en médecine de précision. Nous pensons donc qu'il serait raisonnable d'identifier des biomarqueurs de réponse qui est reproductible pour les inhibiteurs de l'axe BRAF/MEK. Cette observation a été faite dans le chapitre 2 lors de la comparaison des associations voie de signalisations-médicament « pathway-drug associations » qui

montraient une meilleure corrélation entre les études CCLE et GDSC pour les inhibiteurs de MEK.

Nos résultats montrent que l'intégration de plusieurs types de données pour les médicaments pourra éclairer les différents mécanismes moléculaires responsables de la réponse favorable/résistance au médicament. La taxonomie DNF englobant l'ensemble de données NCI60 a également identifié un certain nombre de communautés bien caractérisées. Ceux-ci comprennent la communauté composée d'inhibiteurs de l'EGFR. Nos résultats pour les glycosides cardiaques sont concordants également avec l'étude de Khan et al (Khan et al., 2014). Ces composés inhibent les pompes Na +/K + dans les cellules. En utilisant une approche, basée sur les pharmacophores 3D, combinées avec des caractéristiques génomiques, Khan et al, avait également identifié bisacodyl, un médicament laxatif, comme partageant un mécanisme similaire avec les glycosides cardiaques, en dépit de sa divergence structurelle. Notamment, notre taxonomie intégrative récapitule ces résultats, ce qui démontre que la combinaison de l'information structurelle et génomique est aussi une stratégie prometteuse pour élucider les mécanismes du médicament.

Nous avons ensuite constaté la présence de communautés de médicaments, étroitement liés, qui ont montré une activité antitumorale en générant des espèces réactives de l'oxygène (elesclomol, fenretinide, l'acide éthacrynique, la curcumine, bortézomib, ménadione, celastrol, withaferin A, parthénolide, thapsigargine). L'acide éthacrynique, un médicament indiqué pour l'hypertension, cluster avec la curcumine, un composé naturel utilisé depuis l'antiquité. L'acide éthacrynique inhibe la glutathion-S-transférase (GSTP1), et induit l'apoptose en générant des espèces réactives de l'oxygène (ROS) (R. Wang et al., 2012). La curcumine a aussi montré une activité antitumorale par production de ROS. Ainsi, nous suggérons que GSTP1 pourrait être une cible potentielle de la curcumine.

Notre analyse exploratoire indique la supériorité de DNF envers la classification des médicaments. Dans l'ensemble, la taxonomie DNF a produit une classification cohérente des médicaments pour de multiples classes fonctionnelles, à la fois dans CTRPv2 et NCI60. La méthode DNF a le potentiel de servir de moteur pour les futures études portant sur l'inférence de MoA pour les nouveaux composés non caractérisés, ce qui représente un défi majeur dans le développement de médicaments ainsi que pour la médecine de précision. Tous les aspects analytiques de la méthode sont disponibles pour la communauté scientifique.

5.1.3 Modèle toxicogénomique in vitro et prédiction du mécanisme de toxicité des médicaments

Nous avons exploité les informations toxicogénomiques générées par le projet TG-GATES, à partir d'échantillons de foie de rats traités avec différents produits chimiques et des hépatocytes exposés aux mêmes composés in vitro. À ce jour, plusieurs études ont utilisé TG-GATES pour construire des prédicateurs toxicologiques pertinents. Par exemple, Zhang et al, ont récemment utilisé ces données pour construire une signature génétique prédictive de l'hépatotoxicité et la néphrotoxicité (Zhang et al. 2014). En effet, cette étude a révélé l'importance des gènes de réponse précoce dans le déclenchement de réseaux de signalisation associés à la toxicité. Cette signature prédictive est dérivée d'une période de moins de 24 heures de traitement.

Nous avons développé un pipeline original pour analyser l'ensemble de données TG-GATES en comparant les changements fonctionnels, sous la forme de réponses transcriptionnelles, qui sont induites par une gamme variée de produits chimiques, incluant des médicaments connus, *in vivo* (foie de rat), *in vitro* (hépatocytes primaires en culture) et interespèces (rat vs humain). Une caractéristique principale de notre approche réside dans le fait qu'elle consiste à utiliser les processus biologiques et voies de signalisation qui permettent de tester l'enrichissement de certaines voies affectées par les médicament. Dans ce cas la comparaison se fait entre les espèces sans avoir à considérer seulement les gènes orthologues ce qui compliquerait l'analyse et l'interprétation en aval tel que présenté dans l'étude de Iskar et al. (Iskar et al. 2013). En effet, notre approche a permis une exploration complète des ensembles de données TG-GATES et l'identification des voies fonctionnelles modifiées par les traitements chimiques à la fois chez le rat et humain.

Nos résultats indiquent que la réponse des hépatocytes aux contraintes chimiques peut être analogue *in vitro* et *in vivo*. Plus précisément, nous avons identifié treize modules

hautement conservés représentatifs de la réponse précoce des hépatocytes à l'exposition chimique. Deux d'entre eux sont associés au développement du cancer du foie, à savoir le module de la superfamille des récepteurs TGF— ßR et NOTCH. Étant donné le rôle que le TGF— ßR et les voies NOTCH jouent en réponse à la toxicité précoce (Zhang et al. 2014) et dans le maintien des fonctions hépatiques normales (Morell et Strazzabosco 2014), respectivement. Il ne fut pas surprenant que ces modules aient été surtout affectés par des hépatocarcinogènes et produits toxiques environnementaux. Ce qui fut moins attendu, d'après nos résultats, est le fait que ces deux voies sont significativement associées aux hépatocarcinogènes seulement chez l'homme et non chez le rat. Cela peut refléter une différence fondamentale dans la façon dont les deux espèces traitent ces produits chimiques.

Nous avons aussi montré que les agonistes de PPARa (clofibrate, fénofibrate, gemfibrozil, benziodarone et benzbromarone) sont associés significativement à l'activation de PPAR-alpha uniquement dans le foie de rat, mais pas dans les hepotacites humains. Ceci pourrait expliquer pourquoi ces composés sont hépatocancérigènes chez les rats, mais pas chez l'homme (Lai, 2004). Notre approche montre qu'il sera possible de capter des signaux d'alarme pour certains médicaments même avant de lancer l'évaluation du potentiel cancérigène dans les phases précliniques.

Nous avons aussi confirmé la pertinence biologique de nos biclusters par rapport à une récente étude (Grinberg et al. 2014). En effet, nous avons montré que nos modules récapitulent les réponses dites stéréotypiques, ainsi que des perturbations spécifiques suite à l'exposition aux médicaments. De plus, nous avons suggéré que le module associé au récepteur de TGF—β, en plus d'être enrichi pour des hépatocarcinogènes connus, pourrait agir comme un biomarqueur potentiel de toxicité chimique dans les hépatocytes humains. Il est important de noter que notre nouveau pipeline bioinformatique complète les approches précédentes, utilisées pour élucider les mécanismes de la toxicité chimique *in vitro* ou *in vivo*, en permettant l'exploration efficace des changements moléculaires induits par les produits chimiques. Les modules qui ont émergé de notre analyse globale suggèrent que des réseaux fonctionnels de réponse aux xénobiotiques sont hautement conservés dans le système hépatique entre l'humain et le rat. Toutefois, il faut comprendre qu'il existe des différences dans le métabolisme des médicaments par le biais des enzymes métaboliques entre les systèmes, ce qui complique l'interprétation des données *in vivo* par rapport aux données *in*

vitro. Il faut aussi tenir compte de la variabilité génétique interindividuelle, surtout chez l'homme, dans la réponse aux médicaments. Nous sommes aussi conscients que les annotations dans les bases de données sont incomplètes et peuvent donc limiter cette approche dans une certaine mesure.

Notre étude toxicogénomique clôture la série d'études que nous avons entamée pour comprendre pourquoi le développement préclinique des médicaments fait face à plusieurs défis et certainement plusieurs des médicaments ne seront jamais commercialisés suite a des problèmes d'efficacité et de toxicité. Depuis le chapitre 2 jusqu'au chapitre 4 nous avons tenté d'exploiter les plus grandes bases de données pharmaco et toxicogénomiques pour évaluer la pertinence de plusieurs types de jeu de données, et ceci pour la découverte de biomarqueurs de réponse et la compréhension des mécanismes moléculaires affectées par les médicaments et produits chimiques. Cependant, bien que nos résultats ont permis des avancées substantielles en pharmaco et toxicogénomiques, ils ne constituent qu'une partie du puzzle qu'est le développement de médicaments efficaces dans le cadre de la médecine de précision.

5.1.4 Innovation et impact scientifiques

1— Dans le second chapitre de ma thèse, j'ai développé un pipeline bioinformatique pour étudier les facteurs d'(in) consistance entre deux grandes études pharmacogénomiques (CCLE et GDSC) et suggéré que les différences observées émergeaient plutôt de la non-standardisation des mesures pharmacologiques dans les lignées cellulaires, ce qui pourrait limiter la validation de biomarqueurs de réponse aux médicaments antitumoraux dans les études translationnelles. L'approche comparative que j'ai implémentée fut la première de ce type en pharmacologique. Les résultats obtenus sont importants, car les biomarqueurs sont devenus, de plus en plus, des outils importants dans le développement de thérapies ciblées, l'un des piliers de la médecine de précision.

2— Dans le troisième chapitre de ma thèse, j'ai implémenté un pipeline bioinformatique qui montre la supériorité de l'approche intégrative qui tient en compte diffèrent paramètres pour les médicaments (structure, cytotoxicité, perturbation du transcriptome) pour élucider leur mécanisme d'action (MoA). Notre approche peut être appliquée à plusieurs classes de médicaments puisque les technologies de haut débit (micropuces, criblage pharmacologique) sont disponibles. Cette approche tient en considération les mécanismes moléculaires induits

par le médicament en contraste avec les approches classiques qui classifient les médicaments par similarité chimique/indication thérapeutique. Notre approche présente des perspectives larges pour le repositionnement de médicaments déjà en développement clinique ce qui pourrait conduire à une nouvelle utilisation clinique du médicament.
3— Dans le quatrième chapitre de ma thèse, j'ai développé un pipeline bioinformatique pour étudier le niveau de conservation des mécanismes moléculaires entre les études toxicogénomiques *in vivo* et *in vitro* démontrant que les hépatocytes humains sont des modèles fiables pour détecter les produits toxiques hépatocarcinogènes. Notre approche a généré un répertoire unique de mécanismes moléculaires pouvant expliquer pourquoi certains médicaments induisent des lésions hépatiques graves chez les patients. Puisque la toxicité hépatique est l'une des principales causes du retrait du médicament, notre pipeline pourra être utilisé pour comparer de nouvelles molécules chimiques, en phase préclinique, aux profils toxicogénomiques de notre répertoire.

5.1.5 Validations biologiques et travaux futurs

Bien que nos analyses bioinformatiques aient permis de générer de nouvelles hypothèses pertinentes, il faudra du travail supplémentaire en laboratoire pour valider ces résultats. Certaines des expériences, envisagées dans le laboratoire, sont discutées ci-dessous. De plus, je présenterai les aspects de mon travail qui seront utile pour des projets futurs en pharmacogénomique.

1- **Création d'un grand répertoire qui intègre plusieurs types de jeux de données pharmacogénomiques**

De multiples études pharmacogénomiques fournissent maintenant une unique opportunité pour enquêter de nouvelles thérapies contre le cancer. Comme présenté dans ma thèse, on a collectionné un grand ensemble de jeux de données, ce qui fournit un riche répertoire de données pharmacogénomiques qui permettra d'illuminer sur les voies moléculaires inférant une sensibilité/résistance aux médicaments. À l'heure actuelle, l'utilisation de différentes nomenclatures pour annoter les médicaments et les lignées cellulaires est l'une des limitations pour intégrer ces jeux de données, ce qui pourra retarder leur réutilisation, par exemple pour la méta-analyse. Ceci détient un potentiel sans précédent pour l'identification de biomarqueurs

robustes de la réponse aux médicaments avec applications cliniques potentielles. Pour relever ce défi, nous avons développé *PharmacoGx*, un programme en langage R, qui sera suivi prochainement de *PharmacoDB*, une base de données unifiée, qui est en cours de développement dans notre laboratoire et qui comprend la plupart des jeux de données citées dans ma thèse. Une interface graphique, facile à utiliser, permettra à cette base de données d'être interrogée par des utilisateurs académiques, intéressés par le développement de biomarqueurs ou la compréhension des mécanismes des médicaments. PharmacoDB fournira des annotations détaillées (essais pharmacologiques, mutations associées, profils d'expression…) par médicament ou par lignée cellulaire. En tant que tel, *PharmacoDB* sera une ressource précieuse qui fournira un accès rapide pour tester plusieurs hypothèses biologiques.

2- Études intégratives, mécanisme d'action et validation biologiques

La tendance générale, à partir de notre étude, est que les composés structurellement similaires ont tendance à avoir des profils transcriptomiques similaires indépendamment de l'état biologique étudié. Cependant, certains médicaments structurellement dissemblables partagent les profils transcriptomiques similaires. Cela signifie qu'elles peuvent donner lieu à un effet physiologique similaire, malgré les différences entre les structures chimiques; par conséquent, un médicament pourrait même être utilisé comme agent thérapeutique pour la même indication que l'autre médicament. Par exemple, le disulfiram, un médicament utilisé pour traiter l'alcoolisme chronique, présentait dans la base de données CMAP un profil transcriptomique similaire à des composés chimiques induisant un stress oxydatif dans les cellules cancéreuses. Des validations biologiques menées dans le laboratoire du Dr Jacques Archambault à l'IRCM ont montré qu'évidemment l'induction du stress oxydatif était l'un des MoA du disulfiram dans les cellules cancéreuses testées au laboratoire. Ce qui prouve que la compréhension de la complexité du MoA des médicaments nécessite l'utilisation des données de pharmacogénomiques, qui offrent un aperçu sans précèdent pour disséquer les voies de signalisation moléculaire affectée par les médicaments.

En plus des jeux de données citées dans ma thèse, la présence de large base de données de bioessais tels que PubChem offrira aussi une opportunité unique pour intégrer des données d'activité du médicament contre une ou plusieurs cibles moléculaires.

On ajoutera aussi des jeux de données de toxicogénomique pour le médicament à partir des bases de données DRUGMATRIX /TGGATES. Il sera intéressant d'intégrer, pour un médicament donné, des bioessais pharmacologiques et des données de transcriptomique dans des hépatocytes, ce qui pourra illuminer sur le MoA de toxicité du médicament et ainsi grouper des classes de médicaments selon la similarité du MoA dans les hépatocytes. Par exemple, si un médicament, en cours d'investigation préclinique, présente des caractéristiques similaires à certains médicaments hépatotoxiques, il serait raisonnable de considérer le principe de « guilt by association » même dans des phases précoces et essayer de développer des dérivés moins toxiques ce qui pourra freiner l'attrition de plusieurs médicaments et minimiser les pertes économiques. Il est établi que les études de toxicité chez le rat génèrent un grand nombre de faux positifs (le cas de l'hépatocarcinogénèse lié au module de GPCR par exemple, chapitre 4) d'où l'utilité de trouver un modèle qui récapitule la physiologie chez l'humain. Il serait intéressant par exemple de valider expérimentalement, le module du TGFbeta qui est associé à des médicaments et produits chimiques cancérigènes chez le rat et l'Humain.

On développera prochainement une interface graphique (Web-DNF), facile à utiliser, qui permettra d'accéder aux données intégratives de plusieurs centaines de médicaments et ainsi explorer et générer des hypothèses sur le MoA.

3- Des lignées cellulaires à l'humain, défi de l'approche translationnelle

L'un des buts ultimes de la médecine de précision est de pouvoir orienter les options thérapeutiques selon le profil génomique de la tumeur chez le patient. Cependant, notre connaissance de la biologie des tumeurs et le développement de nouvelles thérapies sont entravés par le manque de modèles expérimentaux pertinent. Dans ma thèse, j'ai montré une partie des avantages et des limitations du modèle de lignée cellulaire. Évidemment, en plus des données de microarrays, il serait intéressant d'exploiter les données de séquençage de nouvelle

génération (e.g. RNA-seq) dans des tissus tumoraux de patients pour essayer de corréler ceux-ci avec des signatures transcriptomique dans des lignées cellulaires susceptibles à certains médicaments anticancéreux y compris ceux présents dans notre banque de données. Je propose aussi d'utiliser les technologies récentes tel que CRISPR pour introduire des mutations dans des cellules souches humaines et donc créer un modèle cellulaire qui récapitule les lésions oncogéniques observées dans plusieurs tumeurs agressives telles que le médulloblastome. Dans ce cas, on pourra développer des stratégies thérapeutiques ciblant certains sous types de tumeurs rares. Les méthodes présentées dans ma thèse pourront aussi prioritiser les traitements en sélectionnant des classes de médicaments présentant des caractéristiques moléculaires, pharmacologiques et toxicologiques similaires.

5.2 Conclusion

Les études présentées dans cette thèse ont permis d'explorer l'utilité des jeux de données pharmaco-toxicogénomiques publiées. Dans le chapitre 2, nos approches analytiques ont mis en lumière les sources potentielles d'inconsistance qui pourrait limiter l'utilisation de ces données pour identifier des biomarqueurs de réponse aux médicaments dans le chapitre 3, nous avons montré que l'intégration de plusieurs types de données associées aux médicaments présente un avantage majeur pour élucider la complexité de leur MoA. Et enfin dans le chapitre 4, nous avons montré que dans des hépatocytes humains en cultures, la toxicogénomique a permis d'identifier plusieurs voies de signalisation, affectées par certaines classes de médicaments, et qui sont importantes pour le bon fonctionnement du système hépatique. Dans le futur, je souhaite pouvoir continuer à utiliser des approches semblables pour y intégrer plusieurs types de données pharmacologiques, moléculaires et de données cliniques dans le but de mieux comprendre les bases moléculaires des maladies humaines et identifier de nouvelles approches thérapeutiques.

Bibliographie

1. Kim RS, Goossens N, Hoshida Y. Use of big data in drug development for precision medicine. Expert Rev Precis Med Drug Dev. 2016;1: 245–253.

2. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483: 603–607.

3. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature. 2012;483: 570–575.

4. Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, Ferriero R, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. Proc Natl Acad Sci U S A. 2010;107: 14621–14626.

5. Chen B, Greenside P, Paik H, Sirota M, Hadley D, Butte AJ. Relating Chemical Structure to Cellular Response: An Integrative Analysis of Gene Expression, Bioactivity, and Structural Data Across 11,000 Compounds. CPT Pharmacometrics Syst Pharmacol. 2015;4: 576–584.

6. Grinberg M, Stöber RM, Edlund K, Rempel E, Godoy P, Reif R, et al. Toxicogenomics directory of chemically exposed human hepatocytes. Arch Toxicol. 2014;88: 2261–2287.

7. Sotiriou C, Pusztai L. Gene-expression signatures in breast cancer. N Engl J Med. 2009;360: 790–800.

8. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144: 646–674.

9. Clarke PA, te Poele R, Wooster R, Workman P. Gene expression microarray analysis in cancer biology, pharmacology, and drug development: progress and potential. Biochem Pharmacol. 2001;62: 1311–1336.

10. Alberts B, Johnson A, Lewis J, Morgan D, Raff M, Roberts K, et al. Molecular Biology of the Cell, Sixth Edition. Garland Science; 2014.

11. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012;13: 341.

12. Pollack A. DNA sequencing caught in deluge of data. NY Times . 2011;1. Available: http://beacon-center.org/wp-content/uploads/2010/10/NYT113011_DNASeqDelugeData.pdf

13. O'Driscoll A, Daugelaite J, Sleator RD. "Big data", Hadoop and cloud computing in genomics. J Biomed Inform. 2013;46: 774–781.

14. Krampis K, Booth T, Chapman B, Tiwari B, Bicak M, Field D, et al. Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community. BMC Bioinformatics. 2012;13: 42.

15. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, et al. Galaxy: a platform for interactive large-scale genome analysis. Genome Res. 2005;15: 1451–1455.

16. Clarke L, Zheng-Bradley X, Smith R, Kulesha E, Xiao C, Toneva I, et al. The 1000 Genomes Project: data management and community access. Nat Methods. Nature Publishing Group; 2012;9: 459–462.

17. Schadt EE. The changing privacy landscape in the era of big data. Mol Syst Biol. 2012;8: 612.

18. Luo J, Wu M, Gopukumar D, Zhao Y. Big Data Application in Biomedical Research and Health Care: A Literature Review. Biomed Inform Insights. 2016;8: 1–10.

19. Dalma-Weiszhausz DD, Warrington J, Tanimoto EY, Miyada CG. [1] The Affymetrix GeneChip® Platform: An Overview. Methods in Enzymology. Academic Press; 2006. pp. 3–28.

20. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 2003;31: e15.

21. Mei R, Hubbell E, Bekiranov S, Mittmann M, Christians FC, Shen M-M, et al. Probe selection for high-density oligonucleotide arrays. Proc Natl Acad Sci U S A. 2003;100: 11237–11242.

22. Kohane IS, Butte AJ, Kho A. Microarrays for an Integrative Genomics. Cambridge, MA, USA: MIT Press; 2002.

23. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, et al. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res. 2005;33: e175.

24. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4: 249–264.

25. Hubbell E, Liu W-M, Mei R. Robust estimators for expression analysis. Bioinformatics. 2002;18: 1585–1592.

26. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4: 249–264.

27. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3: Article3.

28. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Series B Stat Methodol. [Royal Statistical Society, Wiley]; 1995;57: 289–300.

29. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005;102: 15545–15550.

30. Klijn C, Durinck S, Stawiski EW, Haverty PM, Jiang Z, Liu H, et al. A comprehensive transcriptional portrait of human cancer cell lines. Nat Biotechnol. 2015;33: 306–312.

31. Wang C, Gong B, Bushel PR, Thierry-Mieg J, Thierry-Mieg D, Xu J, et al. The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. Nat Biotechnol. 2014;32: 926–932.

32. Thompson JA, Tan J, Greene CS. Cross-platform normalization of microarray and RNA-seq data for machine learning applications. PeerJ. 2016;4: e1621.

33. Goodspeed A, Heiser LM, Gray JW, Costello JC. Tumor-Derived Cell Lines as Molecular Models of Cancer Pharmacogenomics. Mol Cancer Res. 2016;14: 3–13.

34. Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. Nat Rev Cancer. 2006;6: 813–823.

35. Shankavaram UT, Varma S, Kane D, Sunshine M, Chary KK, Reinhold WC, et al. CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. BMC Genomics. 2009;10: 277.

36. Heiser LM, Sadanandam A, Kuo W-L, Benz SC, Goldstein TC, Ng S, et al. Subtype and pathway specific responses to anticancer compounds in breast cancer. Proc Natl Acad Sci U S A. 2012;109: 2724–2729.

37. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2013;41: D955–61.

38. Basu A, Bodycombe NE, Cheah JH, Price EV, Liu K, Schaefer GI, et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. Cell. 2013;154: 1151–1161.

39. Reinhold WC, Sunshine M, Liu H, Varma S, Kohn KW, Morris J, et al. CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. Cancer Res. 2012;72: 3499–3511.

40. Bussey KJ, Chin K, Lababidi S, Reimers M, Reinhold WC, Kuo W-L, et al. Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel. Mol Cancer Ther. 2006;5: 853–867.

41. Varma S, Pommier Y, Sunshine M, Weinstein JN, Reinhold WC. High resolution copy number variation data in the NCI-60 cancer cell lines from whole genome microarrays accessible through CellMiner. PLoS One. 2014;9: e92047.

42. MacConaill LE, Garraway LA. Clinical Implications of the Cancer Genome. J Clin Oncol. 2010;28: 5219–5228.

43. Basu A, Bodycombe NE, Cheah JH, Price EV, Liu K, Schaefer GI, et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. Cell. 2013;154: 1151–1161.

44. Seashore-Ludlow B, Rees MG, Cheah JH, Cokol M, Price EV, Coletti ME, et al. Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. Cancer Discov. 2015;5: 1210–1223.

45. Rees MG, Seashore-Ludlow B, Cheah JH, Adams DJ, Price EV, Gill S, et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. Nat Chem Biol. 2016;12: 109–116.

46. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, et al. A gene expression database for the molecular pharmacology of cancer. Nat Genet. 2000;24: 236–244.

47. Lorenzi PL, Llamas J, Gunsior M, Ozbun L, Reinhold WC, Varma S, et al. Asparagine synthetase is a predictive biomarker of L-asparaginase activity in ovarian cancer cell lines. Mol Cancer Ther. 2008;7: 3123–3128.

48. Lorenzi PL, Reinhold WC, Rudelius M, Gunsior M, Shankavaram U, Bussey KJ, et al. Asparagine synthetase as a causal, predictive biomarker for L-asparaginase activity in ovarian cancer cells. Mol Cancer Ther. 2006;5: 2613–2623.

49. Ikediobi ON, Davies H, Bignell G, Edkins S, Stevens C, O'Meara S, et al. Mutation analysis of 24 known cancer genes in the NCI-60 cell line set. Mol Cancer Ther. 2006;5: 2606–2612.

50. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol. Blackwell Publishing Ltd; 2005;67: 301–320.

51. Kern SE. Why your new cancer biomarker may never work: recurrent patterns and remarkable diversity in biomarker failures. Cancer Res. 2012;72: 6097–6101.

52. Papillon-Cavanagh S, De Jay N, Hachem N, Olsen C, Bontempi G, Aerts HJWL, et al. Comparison and validation of genomic predictors for anticancer drug sensitivity. J Am Med Inform Assoc. 2013;20: 597–602.

53. Papillon-Cavanagh S, De Jay N, Hachem N, Olsen C, Bontempi G, Aerts HJWL, et al. Comparison and validation of genomic predictors for anticancer drug sensitivity. J Am Med Inform Assoc. 2013;20: 597–602.

54. Jang IS, Neto EC, Guinney J, Friend SH, Margolin AA. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. Pac Symp Biocomput. 2014; 63–74.

55. Dong S, Kong J, Kong F, Kong J, Gao J, Ji L, et al. Sorafenib suppresses the epithelial-mesenchymal transition of hepatocellular carcinoma cells after insufficient radiofrequency ablation. BMC Cancer. 2015;15: 939.

56. Cortes-Ciriano I, van Westen GJP, Murrell DS, Lenselink EB, Bender A, Malliavin TE. Applications of proteochemometrics - from species extrapolation to cell line sensitivity modelling. BMC Bioinformatics. 2015;16: 1–2.

57. Booth B, Zemmel R. Prospects for productivity. Nat Rev Drug Discov. 2004;3: 451–456.

58. Dimasi JA. New drug development in the United States from 1963 to 1999. Clin Pharmacol Ther. 2001;69: 286–296.

59. Adams CP, Brantner VV. Estimating the cost of new drug development: is it really 802 million dollars? Health Aff . 2006;25: 420–428.

60. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov. 2004;3: 673–683.

61. Terstappen GC, Schlüpen C, Raggiaschi R, Gaviraghi G. Target deconvolution strategies in drug discovery. Nat Rev Drug Discov. 2007;6: 891–903.

62. Ambesi-Impiombato A, Bernardo D d. Computational Biology and Drug Discovery: From Single-Target to Network Drugs. Curr Bioinform. 2006;1: 3–13.

63. Berger SI, Iyengar R. Network analyses in systems pharmacology. Bioinformatics. 2009;25: 2466–2472.

64. Mani KM, Lefebvre C, Wang K, Lim WK, Basso K, Dalla-Favera R, et al. A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. Mol Syst Biol. EMBO Press; 2008;4: 169.

65. Eckert H, Bajorath J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. Drug Discov Today. 2007;12: 225–233.

66. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, et al. Predicting new molecular targets for known drugs. Nature. 2009;462: 175–181.

67. Fourches D, Muratov E, Tropsha A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. J Chem Inf Model. 2010;50: 1189–1204.

68. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006;313: 1929–1935.

69. Barrett T, Edgar R. [19] Gene Expression Omnibus: Microarray Data Storage, Submission, Retrieval, and Analysis. Methods in Enzymology. Academic Press; 2006. pp. 352–369.

70. Peck D, Crawford ED, Ross KN, Stegmaier K, Golub TR, Lamb J. A method for high-throughput gene expression signature analysis. Genome Biol. 2006;7: R61.

71. Roth WD, Wayne D. Personal flow cytometers----luminex. The microflow cytometer Pan Stanford Publishing, Singapore. 2010; 37–50.

72. Duan Q, Flynn C, Niepel M, Hafner M, Muhlich JL, Fernandez NF, et al. LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. Nucleic Acids Res. 2014;42: W449–60.

73. Napolitano F, Zhao Y, Moreira VM, Tagliaferri R, Kere J, D'Amato M, et al. Drug repositioning: a machine-learning approach through data integration. J Cheminform. 2013;5: 30.

74. Woo JH, Shimoni Y, Yang WS, Subramaniam P, Iyer A, Nicoletti P, et al. Elucidating Compound Mechanism of Action by Network Perturbation Analysis. Cell. 2015;162: 441–451.

75. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? Nat Rev Drug Discov. 2004;3: 711–715.

76. Giacomini KM, Krauss RM, Roden DM, Eichelbaum M, Hayden MR, Nakamura Y. When good drugs go bad. Nature. 2007;446: 975–977.

77. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov. 2010;9: 203–214.

78. Hartung T. Toxicology for the twenty-first century. Nature. 2009;460: 208–212.

79. Nuwaysir EF, Bittner M, Trent J, Barrett JC, Afshari CA. Microarrays and toxicology: the advent of toxicogenomics. Mol Carcinog. 1999;24: 153–159.

80. Vinken M, Doktorova T, Ellinger-Ziegelbauer H, Ahr H-J, Lock E, Carmichael P, et al. The carcinoGENOMICS project: critical selection of model compounds for the development of omics-based in vitro carcinogenicity screening assays. Mutat Res. 2008;659: 202–210.

81. Fasinu P, Bouic PJ, Rosenkranz B. Liver-based in vitro technologies for drug biotransformation studies - a review. Curr Drug Metab. 2012;13: 215–224.

82. de Graaf IAM, Olinga P, de Jager MH, Merema MT, de Kanter R, van de Kerkhof EG, et al. Preparation and incubation of precision-cut liver and intestinal slices for application in drug metabolism and toxicity studies. Nat Protoc. 2010;5: 1540–1551.

83. Knowles BB, Howe CC, Aden DP. Human hepatocellular carcinoma cell lines secrete the major plasma proteins and hepatitis B surface antigen. Science. 1980;209: 497–499.

84. Morris KM, Aden DP, Knowles BB, Colten HR. Complement biosynthesis by the human hepatoma-derived cell line HepG2. J Clin Invest. 1982;70: 906–913.

85. Aninat C, Piton A, Glaise D, Le Charpentier T, Langouët S, Morel F, et al. Expression of cytochromes P450, conjugating enzymes and nuclear receptors in human hepatoma HepaRG cells. Drug Metab Dispos. 2006;34: 75–83.

86. Cerec V, Glaise D, Garnier D, Morosan S, Turlin B, Drenou B, et al. Transdifferentiation of hepatocyte-like cells from the human hepatoma HepaRG cell line through bipotent progenitor. Hepatology. Wiley Subscription Services, Inc., A Wiley Company; 2007;45: 957–967.

87. Jennen DGJ, Magkoufopoulou C, Ketelslegers HB, van Herwijnen MHM, Kleinjans JCS, van Delft JHM. Comparison of HepG2 and HepaRG by whole-genome gene expression analysis for the purpose of chemical hazard identification. Toxicol Sci. 2010;115: 66–79.

88. van Delft JHM, van Agen E, van Breda SGJ, Herwijnen MH, Staal YCM, Kleinjans JCS. Discrimination of genotoxic from non-genotoxic carcinogens by gene expression profiling. Carcinogenesis. 2004;25: 1265–1276.

89. Westerink WMA, Schoonen WGEJ. Cytochrome P450 enzyme levels in HepG2 cells and cryopreserved primary human hepatocytes and their induction in HepG2 cells. Toxicol In Vitro. 2007;21: 1581–1591.

90. Olsavsky KM, Page JL, Johnson MC, Zarbl H, Strom SC, Omiecinski CJ. Gene expression profiling and differentiation assessment in primary human hepatocyte cultures, established hepatoma cell lines, and human liver tissues. Toxicol Appl Pharmacol. 2007;222: 42–56.

91. Kanebratt KP, Andersson TB. Evaluation of HepaRG cells as an in vitro model for human drug metabolism studies. Drug Metab Dispos. 2008;36: 1444–1452.

92. Lambert CB, Spire C, Renaud M-P, Claude N, Guillouzo A. Reproducible chemical-induced changes in gene expression profiles in human hepatoma HepaRG cells under various experimental conditions. Toxicol In Vitro. 2009;23: 466–475.

93. Hewitt NJ, Lechón MJG, Houston JB, Hallifax D, Brown HS, Maurel P, et al. Primary hepatocytes: current understanding of the regulation of metabolic enzymes and transporter proteins, and pharmaceutical practice for the use of hepatocytes in metabolism, enzyme induction, transporter, clearance, and hepatotoxicity studies. Drug Metab Rev. 2007;39: 159–234.

94. LeCluyse EL, Witek RP, Andersen ME, Powers MJ. Organotypic liver culture models: meeting current challenges in toxicity testing. Crit Rev Toxicol. 2012;42: 501–548.

95. Swift* B, Pfeifer* ND, Brouwer KLR. Sandwich-cultured hepatocytes: an in vitro model to evaluate hepatobiliary transporter-based drug interactions and hepatotoxicity. Drug Metab Rev. 2010;42: 446–471.

96. Michalopoulos GK, Bowen W, Nussler AK, Becich MJ, Howard TA. Comparative analysis of mitogenic and morphogenic effects of HGF and EGF on rat and human hepatocytes maintained in collagen gels. J Cell Physiol. 1993;156: 443–452.

97. LeCluyse EL. Human hepatocyte culture systems for the in vitro evaluation of cytochrome P450 expression and regulation. Eur J Pharm Sci. 2001;13: 343–368.

98. Amacher DE. The primary role of hepatic metabolism in idiosyncratic drug-induced liver injury. Expert Opin Drug Metab Toxicol. 2012;8: 335–347.

99. Li AP. A review of the common properties of drugs with idiosyncratic hepatotoxicity and the "multiple determinant hypothesis" for the manifestation of idiosyncratic drug toxicity. Chem Biol Interact. 2002;142: 7–23.

100. Chen M, Zhang J, Wang Y, Liu Z, Kelly R, Zhou G, et al. The liver toxicity knowledge base: a systems approach to a complex end point. Clin Pharmacol Ther. 2013;93: 409–412.

101. Zhang J, Doshi U, Suzuki A, Chang C-W, Borlak J, Li AP, et al. Evaluation of multiple mechanism-based toxicity endpoints in primary cultured human hepatocytes for

the identification of drugs with clinical hepatotoxicity: Results from 152 marketed drugs with known liver injury profiles. Chem Biol Interact. 2016;255: 3–11.

102. Uehara T, Ono A, Maruyama T, Kato I, Yamada H, Ohno Y, et al. The Japanese toxicogenomics project: application of toxicogenomics. Mol Nutr Food Res. 2010;54: 218–227.

103. Ganter B, Tugendreich S, Pearson CI, Ayanoglu E, Baumhueter S, Bostian KA, et al. Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. J Biotechnol. 2005;119: 219–244.

104. Sarkans U, Parkinson H, Lara GG, Oezcimen A, Sharma A, Abeygunawardena N, et al. The ArrayExpress gene expression database: a software engineering and implementation perspective. Bioinformatics. 2005;21: 1495–1501.

105. Waters M, Boorman G, Bushel P, Cunningham M, Irwin R, Merrick A, et al. Systems toxicology and the Chemical Effects in Biological Systems (CEBS) knowledge base. EHP Toxicogenomics. 2003;111: 15–28.

106. Hayes KR, Vollrath AL, Zastrow GM, McMillan BJ, Craven M, Jovanovich S, et al. EDGE: a centralized resource for the comparison, analysis, and distribution of toxicogenomic information. Mol Pharmacol. 2005;67: 1360–1368.

107. Burgoon LD, Boutros PC, Dere E, Zacharewski TR. dbZach: A MIAME-compliant toxicogenomic supportive relational database. Toxicol Sci. 2006;90: 558–568.

108. Waters MD, Fostel JM. Toxicogenomics and systems toxicology: aims and prospects. Nat Rev Genet. 2004;5: 936–948.

109. Stierum R, Heijne W, Kienhuis A, van Ommen B, Groten J. Toxicogenomics concepts and applications to study hepatic effects of food additives and chemicals. Toxicol Appl Pharmacol. 2005;207: 179–188.

110. Heijne WHM, Stierum RH, Leeman WR, van Ommen B. The introduction of toxicogenomics; potential new markers of hepatotoxicity. Cancer Biomark. 2005;1: 41–57.

111. Currie RA, Orphanides G, Moggs JG. Mapping molecular responses to xenoestrogens through Gene Ontology and pathway analysis of toxicogenomic data. Reprod Toxicol. 2005;20: 433–440.

112. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25: 25–29.

113. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28: 27–30.

114. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, et al. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 2005;33: D428–32.

115. Burczynski ME, McMillian M, Ciervo J, Li L, Parker JB, Dunn RT 2nd, et al. Toxicogenomics-based discrimination of toxic mechanism in HepG2 human hepatoma cells. Toxicol Sci. 2000;58: 399–415.

116. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, et al. Functional discovery via a compendium of expression profiles. Cell. 2000;102: 109–126.

117. Hamadeh HK, Bushel P, Tucker CJ, Martin K, Paules R, Afshari CA. Detection of diluted gene expression alterations using cDNA microarrays. Biotechniques. 2002;32: 322, 324, 326–9.

118. Bergmann S, Ihmels J, Barkai N. Iterative signature algorithm for the analysis of large-scale gene expression data. Phys Rev E Stat Nonlin Soft Matter Phys. 2003;67: 031902.

119. Prelić A, Bleuler S, Zimmermann P, Wille A, Bühlmann P, Gruissem W, et al. A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics. 2006;22: 1122–1129.

120. Csárdi G, Kutalik Z, Bergmann S. Modular analysis of gene expression data with R. Bioinformatics. 2010;26: 1376–1377.

121. Iskar M, Zeller G, Blattmann P, Campillos M, Kuhn M, Kaminska KH, et al. Characterization of drug-induced transcriptional modules: towards drug repositioning and functional understanding. Mol Syst Biol. 2013;9: 662.

122. Zhang JD, Berntenis N, Roth A, Ebeling M. Data mining reveals a network of early-response genes as a consensus signature of drug-induced in vitro and in vivo toxicity. Pharmacogenomics J. 2014;14: 208–216.

123. Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten simple rules for reproducible computational research. PLoS Comput Biol. 2013;9: e1003285.

124. Haverty PM, Lin E, Tan J, Yu Y, Lam B, Lianoglou S, et al. Reproducible pharmacogenomic profiling of cancer cell line panels. Nature. 2016;533: 333–337.

125. Cancer Cell Line Encyclopedia Consortium, Genomics of Drug Sensitivity in Cancer Consortium. Pharmacogenomic agreement between two cancer cell line data sets. Nature. 2015;528: 84–87.

126. Pozdeyev N, Yoo M, Mackie R, Schweppe RE, Tan AC, Haugen BR. Integrating heterogeneous drug sensitivity data from cancer pharmacogenomic studies. Oncotarget. 2016; doi:10.18632/oncotarget.10010

127. Safikhani Z, El-Hachem N, Quevedo R, Smirnov P, Goldenberg A, Juul Birkbak N, et al. Assessment of pharmacogenomic agreement. F1000Research. 2016; doi:10.1101/048470

128. Dong Z, Zhang N, Li C, Wang H, Fang Y, Wang J, et al. Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. BMC Cancer. 2015;15: 489.

129. Cortés-Ciriano I, van Westen GJP, Bouvier G, Nilges M, Overington JP, Bender A, et al. Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel. Bioinformatics. 2015; doi:10.1093/bioinformatics/btv529

130. Safikhani Z, Freeman M, Smirnov P, El-Hachem N, She A, Quevedo R, et al. Revisiting inconsistency in large pharmacogenomic studies [Internet]. bioRxiv. 2015. p. 026153. doi:10.1101/026153

131. Ma'ayan A, Rouillard AD, Clark NR, Wang Z, Duan Q, Kou Y. Lean Big Data integration in systems biology and systems pharmacology. Trends Pharmacol Sci. 2014;35: 450–460.

132. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11: 333–337.

133. Seashore-Ludlow B, Rees MG, Cheah JH, Cokol M, Price EV, Coletti ME, et al. Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. Cancer Discov. 2015;5: 1210–1223.

134. Solit DB, Garraway LA, Pratilas CA, Sawai A, Getz G, Basso A, et al. BRAF mutation predicts sensitivity to MEK inhibition. Nature. 2006;439: 358–362.

135. Katayama R, Aoyama A, Yamori T, Qi J, Oh-hara T, Song Y, et al. Cytotoxic activity of tivantinib (ARQ 197) is not due solely to c-MET inhibition. Cancer Res. 2013;73: 3087–3096.

136. Vogler M, Weber K, Dinsdale D, Schmitz I, Schulze-Osthoff K, Dyer MJS, et al. Different forms of cell death induced by putative BCL2 inhibitors. Cell Death Differ. 2009;16: 1030–1039.

137. Khan SA, Virtanen S, Kallioniemi OP, Wennerberg K, Poso A, Kaski S. Identification of structural features in chemicals associated with cancer drug response: a systematic data-driven analysis. Bioinformatics. 2014;30: i497–504.

138. Wang R, Liu C, Xia L, Zhao G, Gabrilove J, Waxman S, et al. Ethacrynic Acid and a Derivative Enhance Apoptosis in Arsenic Trioxide-Treated Myeloid Leukemia and Lymphoma Cells: The Role of Glutathione S-Transferase P1-1. Clin Cancer Res. 2012;18: 6690–6701.